

10

McGRAW-HILL  
ENCYCLOPEDIA  
OF SCIENCE  
AND  
TECHNOLOGY

PER-PROG







# *McGraw-Hill Encyclopedia*

**McGRAW-HILL BOOK COMPANY**

NEW YORK ST. LOUIS SAN FRANCISCO DALLAS TORONTO LONDON SYDNEY



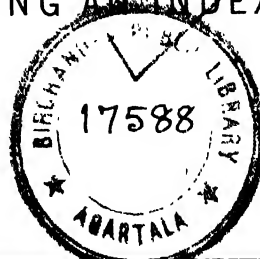
# *of Science and Technology*

AN INTERNATIONAL REFERENCE WORK

IN FIFTEEN VOLUMES INCLUDING AN INDEX

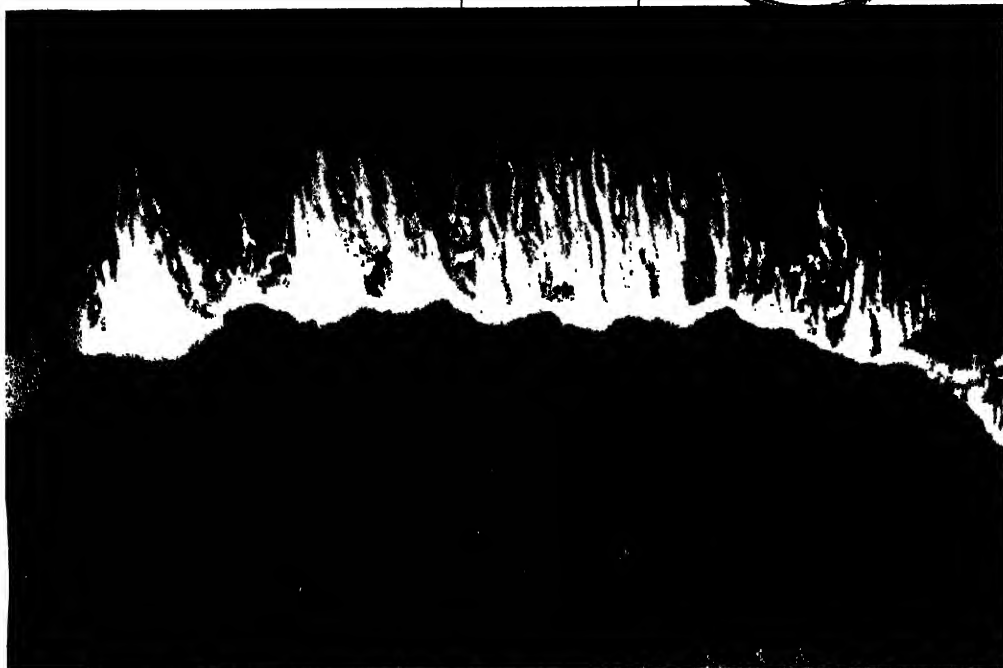
VOLUME 10 PER-PROG

503  
M-478



RETROCONVERTED  
B. C. S. C. L.

26.5 cm.



612-3998  
REFERENCE



# Guide for Readers

## *Basic plan of the encyclopedia*

The subject matter of the various disciplines or branches of science and technology is organized systematically; a general article provides a broad survey of the field, and a number of separate articles, alphabetically arranged, cover its main subdivisions and more specific aspects.

In general, each article begins with a definition of the title that states its scope and coverage. Usually, only the scientific or technological sense is discussed. Most of the articles, after this statement, go on to increasingly complex and detailed considerations. A reader thus needs to proceed only as far as his inclinations and requirements dictate.

Cross references guide the reader from general articles to the other articles into which the subject is subdivided, and from these to articles on more highly specialized phases of the subject. The cross references—there are about 50,000 of them—are printed in capital letters so that they can be easily recognized. By means of the cross references a reader may find his way from ELECTRICAL ENGINEERING, through ELECTRONICS and VACUUM TUBE, to ELECTRON MOTION IN VACUUM or ELECTRON EMISSION. Or, following another line of cross references, the reader would be led to ELECTRIC POWER SYSTEMS, TRANSMISSION LINES, ELECTROMAGNETIC WAVE, and so on.

Every phylum, class, and order in the plant and animal kingdoms is allotted a separate article. Many of the more common families, genera, and species are covered either in one of the order articles or in a separate article under its own scientific or common name.

There are two indexes to information in the encyclopedia, both of them in Volume 15. The comprehensive index, with its 100,000 entries, offers an analytical breakdown: the topical index groups the more than 7200 article titles under nearly 100 general headings, to enable the reader to identify quickly the articles in a subject area.

Most of the longer articles contain bibliographies citing useful sources of further information. For additional bibliographical citations, the reader should refer to related articles (as indicated by the cross

references in the article). Bibliographies are placed at the ends of articles or sometimes at the ends of major sections in long articles.

A list of initials and names of the contributors to the encyclopedia is to be found in Volume 15. This list will permit quick identification of a contributor's initials after an article. Immediately following this list is a second list of encyclopedia contributors with their affiliations and the titles of articles each has written for the encyclopedia.

## *How titles are alphabetized*

Words used as titles are, wherever possible, given in the singular to permit a consistent alphabetic arrangement. Titles are alphabetized by word and not by letter; for example,

**Earth sciences**  
**Earth tides**  
**Earthmover**  
**Earthquake**

A word used as a noun precedes the same word used adjectivally; thus,

**Mercury (element)**  
**Mercury (planet)**  
**Mercury battery**

or

**Circuit, electronic**  
**Circuit breaker**

Hyphenated terms are alphabetized as single words; for example,

**Animal virus**  
**Animal-feed composition**

## *"Electric" and "electrical"*

The adjectives electric and electrical are used in the following senses. Electric—containing, producing, arising from, actuated by, or carrying electricity, or capable of doing so; as, for instance, electric generator, electric motor, electric wiring. Electrical—related to, pertaining to, or associated with electricity, but not having its properties or characteristics; as, for example, electrical code, electrical engineering.





*McGraw-Hill Encyclopedia of Science and Technology*



# PER *Peracarida to Progression (mathematics)*

## Peracarida

A superorder of the class Crustacea, subclass Malacostraca. Common examples of the orders are the opossum shrimps, aquatic sow bugs, and sideswimmers. The Peracarida includes the orders Amphipoda, Cumacea, Isopoda, Mysidacea, Spelaeogriphacea, and Tanaidacea. These orders share a number of characters, the most notable being that the young develop within a thoracic marsupium, which they leave at a late stage of development. The marsupium is formed by from 1-7 membranous oostegites which extend inward from the coxopodites of the thoracic legs. The eggs or developing young lie free in the space between the ventral surface of the thorax and the overlapping oostegites. Members of the order Thermosbaenacea have a dorsal marsupium, formed by the posterior portion of the carapace; consequently they are excluded from the Peracarida by some authorities.

Other features which distinguish the Peracarida from other groups of Malacostraca are the following: The protopodite of the second antenna usually consists of three segments. The thoracic legs are flexed between the fifth and sixth segments. The heart is generally elongate, extending through the greater part of the thoracic region; in the isopods, it may extend into, or lie entirely in, the abdomen, where respiratory exchange occurs. Spermatozoa are usually filiform, in contrast to their spherical or vesicular form in the Eucarida and Hoplocarida.

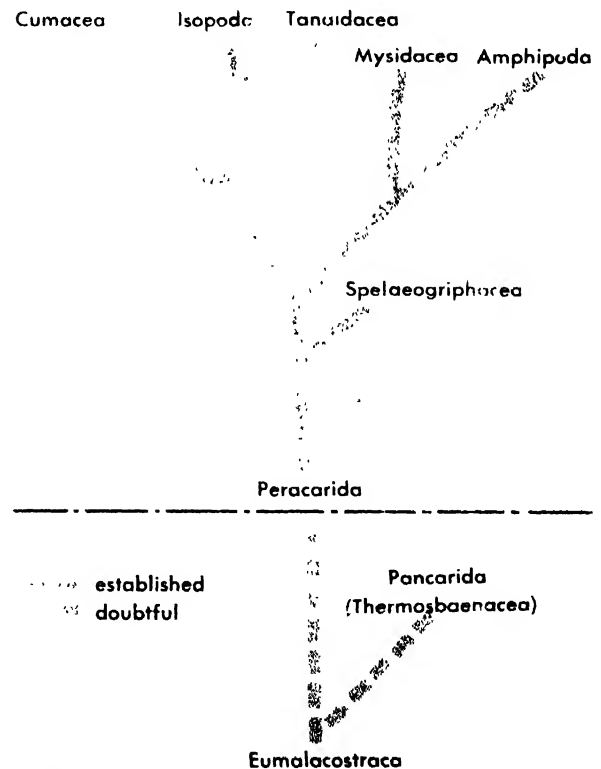
**Morphology.** As in other Malacostraca, the thorax consists of eight somites, the first of which is fused with the head. A carapace is present in all orders but the Amphipoda and Isopoda, but it does not coalesce with more than four thoracic somites as follows: one segment in the Spelaeogriphacea, one or two in the Tanaidacea, one to four in the Mysidacea, and three or four in the Cumacea.

Epipodites are present on the maxillipeds of the orders having a carapace, and their movements keep a current of water flowing along the inner surface of the carapace. In the mysid suborder Mysida, this inner surface is highly vascular, and gaseous exchange takes place through it. Members of the mysid suborder Lophogastrida, however, have epipodial gills on the thoracic legs, and in the cumaceans the maxillipedal epipodite forms a complex gill. The isopods have flattered epipodites on the maxillipeds, but they are apparently not branchial. In the females of some groups of isopods, a lappet of the coxopodite of the maxilliped projects backward and together with the epipodite and an ex-

pansion of the basipodite forms a lamellar plate. Vibratory movements of the maxilliped, including this plate, send a current of water through the marsupium, which aerates the developing embryos. Respiration in isopods takes place through the surface of the pleopods. In the amphipods gaseous exchange occurs in the epipodial gills attached to the inner surfaces of the thoracic legs. There is no epipodite on the maxilliped.

Respiratory epipodites are found only in the Spelaeogriphacea, where they are present on the fifth to seventh thoracic legs.

The thorax is followed by an abdomen of six segments bearing up to five pairs of biramous swimming legs, or pleopods. On the terminal segment there is a pair of uropods which, together with the telson, form a tail-fan. The pleopods are absent in all female and some male cumaceans. In female tanaidaceans they may be reduced or absent, and they may be rudimentary in female mysids. In amphipods the abdominal appendages are sharply divided into two groups; the three anterior pairs are turned forward and are natatory, while the last three, all called uropods, are turned backwards. As



Relationships within the Peracarida.

mentioned before, some or all of the pleopods in the isopods function as gills, and the surfaces may be plicated to increase the respiratory area. One of the anterior pairs of pleopods, or in the suborder Valvifera, the uropods, may be modified into an operculum protecting the delicate respiratory pleopods. In some mysids and isopods one or two pairs of pleopods are modified in the males to assist in the transfer of sperm to the female.

**Relationships among Peracarida.** The Peracarida can be divided into two groups. The probable relationships are shown in the illustration. Some of the characters on which these relationships are based are presented.

The embryo in the Amphipoda is bent ventrad and lies with the dorsal side toward the outside of the egg. The position is reversed in the cumaceans, tanaids, isopods, and mysids. Cumaceans, isopods, and perhaps tanaids leave the egg with the eighth thoracic leg absent. Mysids and amphipods hatch with all their appendages.

Between broods, females of cumaceans and some isopods lose their oostegites. In mysids and amphipods the oostegites are retained during this period.

The lophogastrid mysids have antennal and maxillary nephridia, the latter being small. Cumaceans, tanaids, and isopods have maxillary nephridia, while amphipods and the mysids, other than lophogastrids, have antennal nephridia.

Other criteria, mainly details of the anatomy of the digestive tract, have been treated by R. Siewing.

[T.E.B.]

**Bibliography:** W. Kükenhals and T. Krumbach (eds.), *Handbuch der Zoologie*, vol. 3, no. 6, 1927; R. Lankester (ed.), *A Treatise on Zoology*, pt. VII, fasc. 3, 1909; R. Siewing, Besteht eine engere Verwandtschaft zwischen Isopoden und Amphipoden?, *Zool. Anz.*, 47(7-8):166-180, 1951; R. Siewing, Untersuchungen zur Morphologie der Malacostraca (Crustacea), *Zool. Jahrb. Abt. Anat. u. Ontog. Tiere*, 75:39-176, 1956.

## Perception

The process by which an individual is acquainted with his immediate surroundings. It can be defined by such behavior as looking, listening, and touching, or otherwise reacting with discrimination to the objects and events of the environment. For the human animal it can also be defined in terms of precise verbal activities such as naming, describing, comparing, and distinguishing objects. Finally, inasmuch as it seems to be possible for the human observer to note the process of perception as it occurs in himself, it can be defined as awareness of the external world, or consciousness or experience of it.

However defined, perceiving is distinguished from remembering, which refers to past events, from expecting, which refers to future events, and from imagining, which refers to absent or nonexistent states of affairs. But in no case can these distinctions be sharply drawn, since the "present" environment in time and space cannot in fact be divided

from either the past, the future, or the distant environment by any sharp line of demarcation. "Now" is not a single instant of time, and "here" is not a single point of space. There is a considerable span of perception in time and a considerable range of perception in space. Hence perceiving is not clearly separable from knowing in the general sense of the term. This discussion is concerned primarily with perception and only secondarily with apprehension.

Acquaintance with the environment is obviously dependent on the senses, and therefore the study of the perceptual process is inseparable from the study of the sensory processes (see SENSATION). As John Locke put it, "knowledge comes through the senses and from no other source." To believe, on the contrary, in a mysterious process of intuition or in the efficacy of innate ideas or the power of pure reason has never been popular in Western scientific thought.

Specifically, this means that perceiving depends on the stimulating of receptive mechanisms such as the eye, ear or the skin of an individual by energy in the surrounding environment. The kinds of energy to which human beings and animals are sensitive are light, heat, sound, chemical energy, mechanical force, and gravitational force. The energy at a sense organ is called the proximal stimulus; the source of light, sound, odor, or mechanical pressure is sometimes called the distal, or distant, stimulus. Usually, a distal object or event causes the stimulation of several sense organs at the same time, and the contemporary modes of stimulation combine to form a single percept. A fire is seen, heard, smelled, and its warmth is felt. Usually there is more than enough sensory input to yield a percept. But when stimulation of one channel fails for any reason, perception depends on those remaining; and when stimulation of all channels is eliminated perception ceases.

Perception thus depends on events at several successive stages: the external situation, the energy at the receptors, the excitation of the receptors, the transmission of neural impulses, and complex processes in the brain which make possible both perception and behavior. If the chain of events is interrupted at any stage, the whole channel stops functioning. For example, vision will fail if (1) there is nothing to see, as with an observer floating in empty space, (2) there is no illumination, as in pitch darkness, (3) the light entering an eye is diffused, as with spectacles of ground glass, (4) the light entering an eye is interrupted, as when the eyelids are closed, (5) the eye becomes opaque, as with cataract, (6) the retina is damaged, as in glaucoma, (7) the optic nerve is cut, or (8) the area of the brain to which it leads is destroyed. In any of these cases, a man will be effectively blinded. See VISION.

**As an exploratory process.** Perceiving differs from sensing, as this term is traditionally used, in being an active rather than a passive, or merely receptive, process. Introspectively, it is character-



## Perception

events in the environment	exteroceptive stimuli	motor responses	actions on the environment
pressures at the skin			forces moving the body from place to place (locomotion)
chemicals at nose and mouth	stimulation at the exteroceptive sense organs	motor reactions of the limbs and body	forces moving objects from place to place
sound at the ears			manipulation (tools, etc.)
light at the eyes			

The classical stimulus-response formula. Stimuli produced by responses are left out of account, both those produced by exploratory responses of sense organs

and those produced by gross responses of the musculature. Hence the diagram below should be superposed on this one.

### the modification of stimulation by reactions of the exteroceptive sense organs

exploration with fingers

savoring and sniffing with mouth and nose

cocking head or pricking up ears

focusing, fixating, converging and pursuing with eyes

stimulation at the exteroceptive sense organs

### the modification of reactions by stimulation of the proprioceptive system

intramuscular stimulation

tendon and joint stimulation

tactile stimulation

inner ear stimulation

visual stimulation

motor reactions of limbs and body

The feedback loops for exploring or enhancing external stimulation and those for controlling behavior.

The angular lines represent physical actions; the curved lines represent neural actions.

ized by what is called selective attention, or an effort after clearness, or a search for meaning. Behaviorally, it always involves, at the very least, some kind of adjustment of the sense organs.

For example, the primary reaction of the human eye to light is in focusing the image (accommodation) and centering the fovea on successive details (exploratory fixation). Contact stimulation normally involves movements of the fingers to produce pressures on the skin, an active process of touching rather than a passive one of being touched. Taste and smell involve savoring and sniffing. Even

hearing leads to such listening responses as assist one in localizing the sounding object. The chain of events in a sensory channel, then, involves a circle in the chain, and thereby stimulation leads to adjustments producing more stimulation. There is present, by analogy with electronic systems, a "feedback loop" in every sensory system. It seems to have the function of making stimulation optimal for perception. The focusing of the retinal image registers detail, and the exploration of the ambient light registers a whole set of details. Focusing and exploration are necessary for the accurate percep-

mentioned before, some or all of the pleopods in the isopods function as gills, and the surfaces may be plicated to increase the respiratory area. One of the anterior pairs of pleopods, or in the suborder Valvifera, the uropods, may be modified into an operculum protecting the delicate respiratory pleopods. In some mysids and isopods one or two pairs of pleopods are modified in the males to assist in the transfer of sperm to the female.

**Relationships among Peracarida.** The Peracarida can be divided into two groups. The probable relationships are shown in the illustration. Some of the characters on which these relationships are based are presented.

The embryo in the Amphipoda is bent ventrad and lies with the dorsal side toward the outside of the egg. The position is reversed in the cumaceans, tanaids, isopods, and mysids. Cumaceans, isopods, and perhaps tanaids leave the egg with the eighth thoracic leg absent. Mysids and amphipods hatch with all their appendages.

Between broods, females of cumaceans and some isopods lose their oostegites. In mysids and amphipods the oostegites are retained during this period.

The lophogastrid mysids have antennal and maxillary nephridia, the latter being small. Cumaceans, tanaids, and isopods have maxillary nephridia, while amphipods and the mysids, other than lophogastrids, have antennal nephridia.

Other criteria, mainly details of the anatomy of the digestive tract, have been treated by R. Siewing.

[T.E.B.]

**Bibliography:** W. Kükenthal and T. Krumbach (eds.), *Handbuch der Zoologie*, vol. 3, no. 6. 1927; R. Lankester (ed.), *A Treatise on Zoology*, pt. VII, fasc. 3, 1909; R. Siewing, Besteht eine engere Verwandtschaft zwischen Isopoden und Amphipoden?, *Zool. Anz.*, 47(7-8):166-180, 1951; R. Siewing, Untersuchungen zur Morphologie der Malacostraca (Crustacea), *Zool. Jahrb. Abt. Anat. u. Ontog. Tiere*, 75:39-176, 1956.

## Perception

The process by which an individual is acquainted with his immediate surroundings. It can be defined by such behavior as looking, listening, and touching, or otherwise reacting with discrimination to the objects and events of the environment. For the human animal it can also be defined in terms of precise verbal activities such as naming, describing, comparing, and distinguishing objects. Finally, inasmuch as it seems to be possible for the human observer to note the process of perception as it occurs in himself, it can be defined as awareness of the external world, or consciousness or experience of it.

However defined, perceiving is distinguished from remembering, which refers to past events, from expecting, which refers to future events, and from imagining, which refers to absent or nonexistent states of affairs. But in no case can these distinctions be sharply drawn, since the "present" environment in time and space cannot in fact be divided

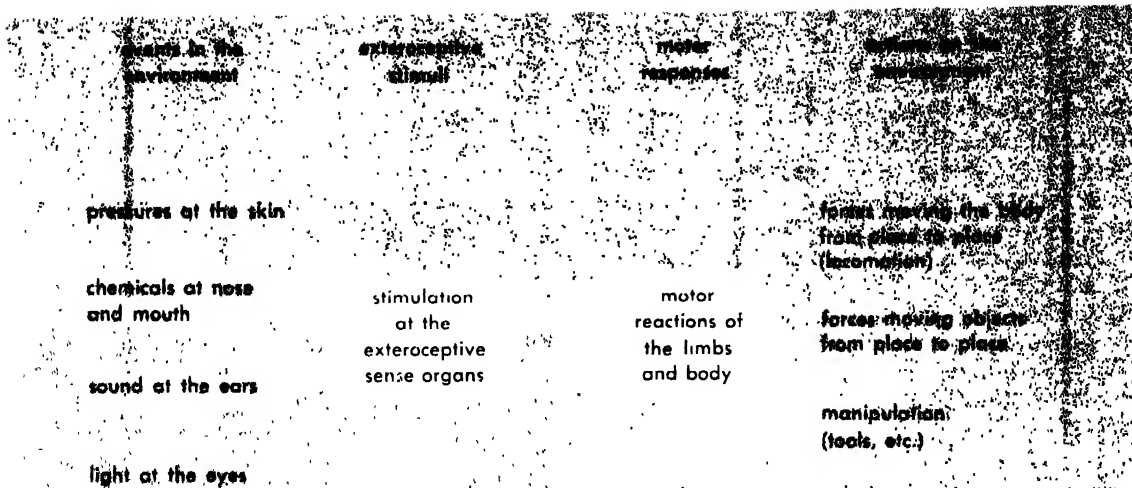
from either the past, the future, or the distant environment by any sharp line of demarcation. "Now" is not a single instant of time, and "here" is not a single point of space. There is a considerable span of perception in time and a considerable range of perception in space. Hence perceiving is not clearly separable from knowing in the general sense of the term. This discussion is concerned primarily with perception and only secondarily with apprehension.

Acquaintance with the environment is obviously dependent on the senses, and therefore the study of the perceptual process is inseparable from the study of the sensory processes (see SENSATION). As John Locke put it, "knowledge comes through the senses and from no other source." To believe, on the contrary, in a mysterious process of intuition or in the efficacy of innate ideas or the power of pure reason has never been popular in Western scientific thought.

Specifically, this means that perceiving depends on the stimulating of receptive mechanisms such as the eye, ear or the skin of an individual by energy in the surrounding environment. The kinds of energy to which human beings and animals are sensitive are light, heat, sound, chemical energy, mechanical force, and gravitational force. The energy at a sense organ is called the proximal stimulus; the source of light, sound, odor, or mechanical pressure is sometimes called the distal, or distant, stimulus. Usually, a distal object or event causes the stimulation of several sense organs at the same time, and the contemporary modes of stimulation combine to form a single percept. A fire is seen, heard, smelled, and its warmth is felt. Usually there is more than enough sensory input to yield a percept. But when stimulation of one channel fails for any reason, perception depends on those remaining; and when stimulation of all channels is eliminated perception ceases.

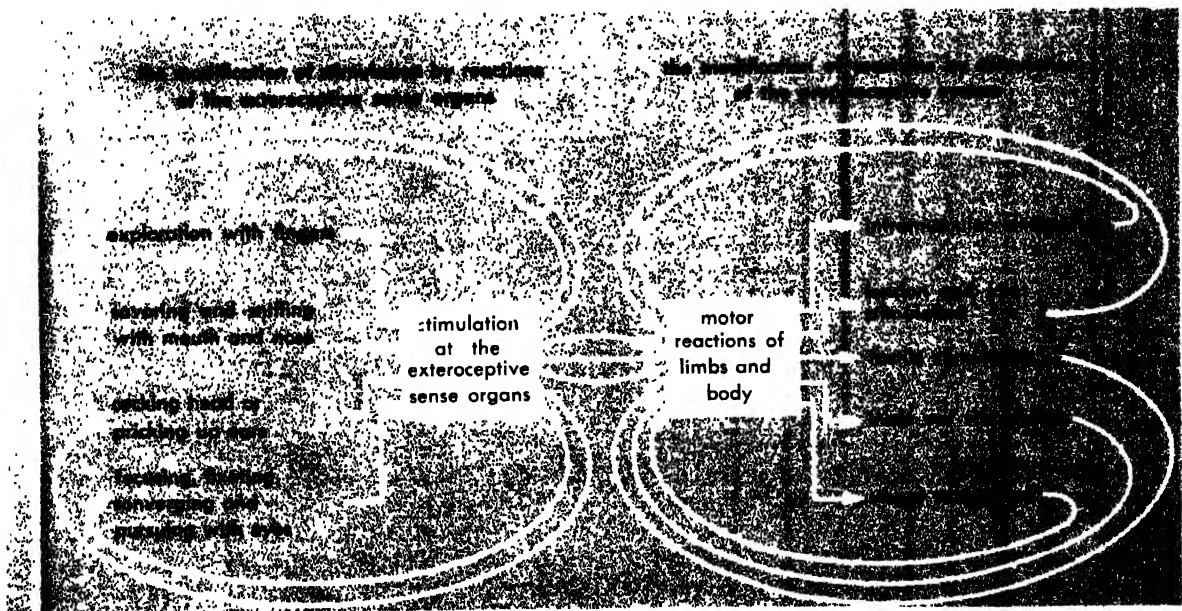
Perception thus depends on events at several successive stages: the external situation, the energy at the receptors, the excitation of the receptors, the transmission of neural impulses, and complex processes in the brain which make possible both perception and behavior. If the chain of events is interrupted at any stage, the whole channel stops functioning. For example, vision will fail if (1) there is nothing to see, as with an observer floating in empty space, (2) there is no illumination, as in pitch darkness, (3) the light entering an eye is diffused, as with spectacles of ground glass, (4) the light entering an eye is interrupted, as when the eyelids are closed, (5) the eye becomes opaque, as with cataract, (6) the retina is damaged, as in glaucoma, (7) the optic nerve is cut, or (8) the area of the brain to which it leads is destroyed. In any of these cases, a man will be effectively blinded. See VISION.

**As an exploratory process.** Perceiving differs from sensing, as this term is traditionally used, in being an active rather than a passive, or merely receptive, process. Introspectively, it is character-



The classical stimulus-response formula. Stimuli produced by responses are left out of account, both those produced by exploratory responses of sense organs

and those produced by gross responses of the musculature. Hence the diagram below should be superposed on this one.



The feedback loops for exploring or enhancing external stimulation and those for controlling behavior.

The angular lines represent physical actions; the curved lines represent neural actions.

ized by what is called selective attention, or an effort after clearness, or a search for meaning. Behaviorally, it always involves, at the very least, some kind of adjustment of the sense organs.

For example, the primary reaction of the human eye to light is in focusing the image (accommodation) and centering the fovea on successive details (exploratory fixation). Contact stimulation normally involves movements of the fingers to produce pressures on the skin, an active process of touching rather than a passive one of being touched. Taste and smell involve savoring and sniffing. Even

hearing leads to such listening responses as assist one in localizing the sounding object. The chain of events in a sensory channel, then, involves a circle in the chain, and thereby stimulation leads to adjustments producing more stimulation. There is present, by analogy with electronic systems, a "feedback loop" in every sensory system. It seems to have the function of making stimulation optimal for perception. The focusing of the retinal image registers detail, and the exploration of the ambient light registers a whole set of details. Focusing and exploration are necessary for the accurate percep-

tion of one's surroundings. The perceiver searches the array of light for meaningful shapes in the overall pattern.

The stimulus for an active sense organ, as distinguished from the stimulus for a receptor cell of the organ, consists of patterns of energy and sequences of pattern. These are the relevant stimuli for perception. As collections of cells, the retina and the cochlea are sensitive to the simple variables of frequency and intensity. The retina is specialized for light energy, and the cochlea for sound energy. But as organs of the body, the eyes and the ears are sensitive to patterns and transitions of frequency and intensity in the array of light and in the flux of sound. The mosaic of receptors in the retina or the cochlea can be excited in enormously complicated combinations of adjacent and successive order. Accordingly, the eye and the ear as total systems can respond to the complex variables of pattern and sequence in the total stimulating situation.

Considering the acts of looking, listening, feeling, smelling, and tasting as selective responses, it becomes evident that there is always more potential stimulus information at the surface of an individual at any given time than he can possibly handle. Moreover, there are always more potential sequences of stimulation than the one or the few he can choose to explore. The environment is to be regarded as an inexhaustible reservoir of potential stimuli. The perceptual process is one of selecting and pursuing those which are important for the individual. Perception is motivated by interest, that is, by vigilant curiosity. Its aim is the registration and clarification of objective facts. But, obviously, different facts are important for different individuals, or for the same individual at different times. Perception depends on what the perceiver is interested in.

**As a process of monitoring behavior.** Although perceiving is traditionally defined as acquainting the individual with his environment, it also acquaints the individual with his own action. It has, in the words of C. S. Sherrington, the famous nineteenth-century physiologist, a proprioceptive function as well as an exteroceptive function whenever the individual is engaged in any kind of behavior beyond mere contemplation of the environment.

In addition to the sensory channels thus far considered, there is the kinesthetic. This is served by various kinds of receptors in the muscles, tendons, and joints, and by still others in the inner ear (see KINESTHETIC SENSATION). Conceived broadly, the sensitivity of an individual to the behavior in which he is engaged is much more than a single sense. It includes all stimulation produced by responses, and this extends at least to touch and vision. When handling things or walking about, a man feels his activity by the changing pressures on the skin, and he sees it by the changing patterns at the eye; he even hears it by the noises at the ear unless he is careful to be silent during manipulation or locomotion. There are many channels for

this "feedback" from response, some of the stimuli being inside the skin of the organism and some being outside, but all of these circular loops have a common function: they keep the individual informed about the outcome of his muscular movements and about the progress of his action. They also serve to guide or control behavior, inasmuch as they define the terminal act of a course of action and indicate the degree to which it has not yet been reached (see CYBERNETICS).

The kind of perception which is produced by action is not as familiar as the kind which is produced by the environment as such. It is harder to investigate, and it is only beginning to be understood, but obviously, the two kinds must be interlocked. In locomotion, for example, an animal needs to perceive both that he is moving through his environment and also where he is going and how close he is to his goal. For this he must be able to perceive the layout of his environment. The visual control of locomotion, therefore, is not separable from the visual perception of space. A walking man has to be able to feel and see himself moving and, at the same time, to feel and see the motionless ground beneath his feet. This necessity raises the question of the so-called "constancy" of environmental perception, which will be discussed later.

**Relationship to the motivating of behavior.** If the perceptual process acquaints the individual with his own actions as well as with the environment, the question may be asked whether it also informs him of his own inner needs and of those outer objects which afford satisfactions of his needs. Does perceiving motivate action as well as being itself a kind of motivated act? Such questions raise difficult theoretical issues. To some extent, an animal or a man can discriminate among, or perceive, his own biological needs (see HUNGER; PAIN, CUTANEOUS; PAIN, DEEP; THIRST).

Likewise, an animal or a man can identify his food, mate, shelter, and other objects of his environment, beneficial or noxious, with considerable success. But the relation of these disparate facts to the motivation of behavior is a controversial question (see MOTIVATION). A man who comes across a rattlesnake in his path perceives and feels and acts all at the same time, and the psychologist's arbitrary separation of these processes fails.

Introspectively, percepts seem to have a subjective reference as well as an objective reference, but some have much more than others. In a painful experience the subjective pain dominates, whereas in a visual experience the object dominates. But even in visual perception there is the implicit sense of "here," and in auditory perception there is the awareness of "now." However objectively oriented one's experience of the world may be, there lurks in the background a variety of tensions, feelings, emotions, the image of the body, and the consciousness of self. Philosophers and psychologists have long recognized the importance of these subjective facts, but they are not easily put to experimental test and so are given little emphasis in scientific psychology.

**Guessing, supposing, or surmising.** Consideration has been made up to this point of the kind of perception characterized by a feeling of certainty and a testable dependence on the stimulation of sense organs. There are many kinds of apprehension, however, which are less certain and less obviously dependent on stimulation. They are not ordinarily called perceiving but guessing, or supposing, or surmising. There is a great difference between observation and inference, but at the same time there is often supposed to be an element of inference in all perception, even the simplest kind. This arises from two theoretical assumptions, first, that only the so-called sensations are certain and clear, and, second, that they are meaningless. On this assumption, some theory of a perceptual process which supplements the sensations is necessary. Consequently there has been a great interest in the various types of uncertain apprehension with the hope that they will reveal the nature of the internal contribution to sensory data.

Many kinds of guessing and its variants may occur, but they are not easily distinguished by introspection. There are experimental methods, however, for producing some of them, and these may be listed as giving a partial classification.

*Careless observation or divided attention.* This may be obtained if the experimenter distracts the subject from the stimulus he is supposed to judge. The accuracy of perception suffers. The experimenter may require the subject to fix his eyes on a point to one side of the presented stimulus so that its image falls on the peripheral retina. Only the most general features of the object are then reportable, and if it is atypical in any respect, the percept will be inexact.

*Strange and unfamiliar things.* An observer who is presented with a wholly novel stimulus cannot identify it. Many foreign words are not heard as being different from one another, and novel patterns do not look different unless they are put close together and compared. Only gross or familiar differences are noted. The novice at using a microscope must learn what to look for and how to perceive the subtleties of color and structure in the slide of tissue under examination. With such stimuli, however, guessing gives way to exact perception as practice in observation continues.

*Obscure or hidden properties of things.* As the medical profession knows, the interpretation of x-ray photographs or the diagnosis of rare diseases is subject to error. In part, this may be due to lack of perceptual skill, but it may also be due to the intrinsic poverty of the information available in sensory stimulation. However actively he searches, the perceiver may not be able to detect valid indicators of the true state of affairs. The same rule may hold for detecting the vintages of wines, or estimating the character of one's acquaintances.

*Experimentally impoverished stimulation.* It has long been known that sensory detection becomes guessing when the absolute energy applied to one of the sensory channels is sufficiently reduced, and

that sensory discrimination becomes guessing when the absolute difference between two energies is reduced. The threshold intensity of light, sound, taste, or touch is defined as that at which the frequency of correct guesses reaches the level of chance. Similarly, the difference threshold is that difference which divides chance guessing from better-than-chance guessing (see SENSATION). Perceptual experience becomes guessing, in the same way, when the pattern in light or sound is impoverished or distorted. The pattern of light can be blurred by optical devices, or veiled by glare, or reduced in size to a threshold level; the pattern of sound can be masked by noise, or disrupted by electronic means, or distorted until speech becomes unintelligible. A commonly used visual device is the tachistoscope, which reduces the duration of the stimulus picture to a fraction of a second or less. Whichever of these methods is used, the result is the same: the poorer the pattern, the less accurate are the reports of the observer.

*No stimulation.* When all channels of stimulation known to the experimenter have been eliminated, guesses can still be made by the subject. They are ordinarily correct only by chance. But certain experiments on the guessing process, using hidden cards and many judgments, sometimes yield a statistically significant departure from chance. This is often taken as evidence for a kind of extrasensory perception, and for an unknown channel of stimulation. See EXTRASENSORY PERCEPTION (ESP). However interesting these results, they should not be considered to apply to perception but simply as an unexplained fact in the study of guessing. They should not be confused with the experiments on weak stimulation, since no variable has been discovered in ESP experiments which correlates with the frequency of correct guesses. For the same reason, as the results now stand, they do not provide evidence for an unknown channel of stimulation.

*Contradictory stimulus information.* Visual patterns can be designed which are ambiguous, that is, which induce either of two perceptions, one incompatible with the other. Painters have manipulated lines and contours in this way, but the Gestalt psychologists were the first to work systematically with ambiguous drawings and to demonstrate the role of contours in shaping phenomenal objects and producing regions which at one time look like filled or solid space and at another time like unfilled empty space. It might seem here that one stimulus paradoxically can arouse two percepts, but it is more likely that two stimulus factors in the same drawing are pitted against one another. To a considerable extent, the observer can select that factor in the total stimulus to which he will respond, or his selection may be influenced by the experimenter.

In all these kinds of apprehension the stimulus applied does not seem to determine the response made, for the latter is subject to other influences. A great variety of such influences has been studied in the effort to get at the nature of the perceptual



process when it is not strictly bound to the stimulus. The influence of verbal suggestion on perception has been demonstrated, as well as an influence of past experience on perception, an influence of emotional prejudice on perception, an influence of cultural background on perception, and the like. But it should be noted that the more demonstrably perception is influenced by these factors the less it is like discerning and the more it is like guessing.

The nature of the subjective component in apprehension, whether the enrichment of stimuli or the organization of stimuli, or something else, is not clear. Theories of the process are a matter of active dispute in psychology. The classical theories of the process have been described by E. G. Boring, the major historian of scientific psychology, and contemporary theories are represented by the authors listed in the bibliography. A consideration of these theories is not given here because such theories can have no firm basis until the extent that ordinary perception is a direct function of the variables of pattern stimulation is known. For that knowledge, other methods of investigation must be used.

#### METHODOLOGY OF PERCEPTION

**Methods of studying sense perception.** The best way to investigate the perceptual process is to induce it artificially, that is, to construct an apparatus which will deliver to an individual the necessary stimulation. The experimenter then has to vary this stimulus in such a way as to obtain from the individual reports, judgments, or discriminative reactions which are a function of the variations. This is the psychophysical procedure by which sensory reactions are studied. See *PSYCHOPHYSICAL METHODS*.

When perceptions are studied, the procedure differs in that the stimulus must be a pattern or sequence of light, sound, pressure, or the like. This requirement puts a strain on the ingenuity of the experimenter, for the systematic variation of pattern or sequence is not easy, and the artificial production of stimuli which imitate those of the natural world is an elaborate and sometimes expensive process. It is not surprising, therefore, that in most of the existing experiments on visual perception the patterns of light are those which can be reflected from line drawings, pictures, or nonsense forms. The experimental work of the Gestalt psychologists is largely based on stimulus material of this sort. There is an increasing tendency, however, for experimenters to manipulate and control the variables of patterned light, as is evidenced by the current interest in optical "texture." For example the perception of a solid surface can be induced, where no solid surface exists, by an optical arrangement in which the light entering the eyes of an observer is patterned with a sufficient degree of density (J. J. Gibson et al., 1955).

Pictures, motion pictures, and binocular stereoscopic motion pictures are ways of approximating the natural input of the eyes to an increasing de-

gree. But the resulting perceptions lack the ultimate degree of "realism" because the pictures usually represent only a small sector of an environment. Methods of reproducing more of the total optic array have been devised, however, in the form of "simulators" for the purpose of military training. The visual input of a helicopter pilot or a tank driver can be imitated by wide-screen projection methods. Moreover, the motions in this field can be made to depend exactly on the control reactions of the driver or the pilot in a "mock-up" of the situation, so that the visual feedback acquaints him with what it is like to control the vehicle. The panoramic motion picture, the Cinerama, provides an optical array which more nearly approaches that of a real environment. Such presentations, if systematically varied and guided by theory, promise to solve many of the problems of visual perception.

For the study of auditory perception the artificial production of pattern and sequence is not so difficult as for visual. The input to the ears is easier to control than the input to the eyes. Recent advances in electronic and mechanical devices enable the experimenter to manipulate air vibrations in the physically complex ways that are characteristic of informative sounds. Not only can natural sounds be recorded and reproduced with fidelity, they can also be produced to order. Speech sounds, for example, can be synthesized. From this it is only a step to the psychophysics of speech perception.

When animals or young children are used as subjects in perception experiments, the experimenter must be content with behavioral discriminations. Verbal discriminations and descriptions by a human adult provide somewhat more direct and subtle evidence of the perceptual process. But the methods are fundamentally the same in that they consist of establishing the dependence of responses on stimuli.

The controllable types of stimulation for the experimental study of perception depend on the kinds of energy to which the external receptors will respond, whether optical, acoustical, thermal, chemical, or mechanical. Examples of uncontrollable stimuli are the force of gravity as registered by the inner ear, which cannot now be experimentally eliminated, and the stimuli originating within the body, which cannot be experimentally controlled. If a pattern and sequence of stimulation which ordinarily arises from a real object or event is applied to the sense organs by an experimenter, the object or event will be perceived as if real, and with all its richness of meaning. If the variables of adjacent and successive order in the stimulus are manipulated, the phenomenal object will vary correspondingly. There is no difference in this respect between the distance receptors and the contact receptors, that is, for example, between the seeing of an object by light at the eye and the feeling of an object by contact at the skin. Both light and mechanical impact ordinarily carry information about solid substances in the environment, albeit somewhat different information. The ability to apprehend

solidity at a distance by means of light is no more a special problem for perception than the ability to apprehend solidity at zero distance by touch, or, conversely, perception of a surface by touch is no more self-explanatory than perceiving it by sight. In either case, some property of the stimulus must specify solidity.

A substance must be in contact with the tongue in order to be tasted; it must be touched so that its texture or temperature may be felt. One must come fairly close to an object in order to smell it, but a large number of its properties can be identified from afar if it can be seen. The properties that enable such identification of an object are color, texture, shape, size, distance, mobility, its place in the layout of surfaces comprising the rest of the environment, and its identity, that is, the properties that distinguish it from other objects. If it is a noisy object, or a being which vocalizes, identification can often be made even in darkness or around a corner. All such discriminations may be tested experimentally. Taken together, they make up the perception of the natural world. The information which the individual obtains concerning the state of the world and his situation in it is conveyed to him wholly by stimulation.

#### PROBLEMS OF SENSORY PERCEPTION

The problems of perception used to be formulated after describing the sensations of color, brightness, pitch, loudness, pressure, warmth, cold, pain, sweet, sour, salt, bitter, and so on. The problems consisted of such questions as how space, time, form, and motion are perceived. There were also the questions of how objects are discerned and how their meanings are perceived. Most puzzling was the problem of explaining the constancy of phenomenal objects in the face of the fact that their retinal images do not remain constant even from one moment to the next. The literature of these problems is extensive (E. G. Boring, 1942; J. J. Gibson, 1950; M. D. Vernon, 1952). An attempt will be made here to reformulate the problems in such a way as to represent modern directions of research. There is, first, the perception of environmental spaces; second, the perception of permanent objects; and third, the perception of changes and events. At a higher level of complexity there is a whole series of problems concerned with the perception of other persons, of their actions and of social stimuli they emit, notably words.

**Environmental spaces.** It is very doubtful that the infant first sees a patchwork of colors in two dimensions and then gradually learns to interpret the patches in relief and at various distances away from him. It is more likely that what the infant sees is an undifferentiated, but by no means a flat picture. The theory that it is a flat picture, however, has been taught for so long and is so much a part of the history of Western thought that it is widely taken for granted. The clues for the interpretation of depths and distances are said to be linear perspective, apparent size, interposition, shadows, and

aerial perspective. These are the monocular signs of distance and together with apparent motion that furnishes the parallax clue, and retinal image disparity, which supplies the binocular clue, form bases for such interpretation. This list is a mixture of facts taken from painting, physics, and physiology.

A simpler question is how terrestrial animals, including men, perceive the terrain. The question can be extended to individual terrain features such as canyons, forests, rooms, corridors, roads, air-lanes, and even the planetary system; each item, however, concerns places or specific environments rather than empty space. The investigator is then led to analyze the array of light projected to a station point in a given environment, then to a traveling station point, and finally to the different station points of two eyes. Variables in the projected light which specify the properties of the environment can thus be discovered, including the recession, the slant, and the discontinuities of surfaces. The information is carried by variables of structure and transformation. Perspective, motion parallax, and binocular disparity can be treated as stimulus variables instead of clues, and the stimulating capacities of these variables can be tested experimentally.

Concurrent gradients of pattern, of motion, and of disparity are characteristic of the light reflected from a continuous surface like the ground. In addition, there is usually the cutaneous stimulation from contact with the ground and always the force of gravity acting on the inner ear and the muscles. If the perception of a ground plane can thus be accounted for, it provides a literal basis for understanding the perception of other special environments and places.

**Surfaces and permanent objects.** Texture or pattern in light specifies some thing or entity in that direction; textureless light specifies nothing in that direction. The former typically comes from the earth; the latter from the sky. Against either background, a coherent pattern with a contour specifies a delimited object in that direction. Stationary objects, together with the ground, constitute the permanent environment. Such objects are of many types: obstacles to locomotion, or goals to be approached; vantage points or shelters; edible substances or dangerous ones. They have to be discriminated and identified by color, texture, shape and size so as to react appropriately to them. The question of how this is possible is asked by psychologists. The projected shape and size of the objects are never the same from one moment to the next as the individual moves about. Moreover, the intensity and color of the light reflected from them does not stay constant from noon to sunset. This is one way of putting the problem of the apparent constancy of phenomenal objects. In actual fact, the projected array from the entire permanent environment, ground and objects alike, undergoes a continuous perspective transformation whenever the perceiver moves. It is a question not only of

why objects stay constant in perception but also of why the ground does.

Although various roundabout answers can be made to this query, the simplest is that there are both variant and invariant properties of optical stimulation under transformation, and that the invariants are the stimuli for the perception of invariant objects, that is, permanent and rigid ones. For example, the kind of texture that a surface possesses remains invariant; rectilinearity remains invariant; and such properties as triangularity or quadrangularity remain invariant. The variant or changing properties of the continuous transformation, on the other hand, are stimuli for the perception of locomotion. On occasion, one may get the impression of a nonrigid or quasi-elastic environment during locomotion, but this is a secondary, not a primary, phenomenon. The above answer would explain how an individual can perceive the permanent layout of the environment and perceive himself moving through it, both at the same time.

The invariants in an array of light which specify the color of a surface despite changes in the illumination of the whole layout of surfaces are not yet fully understood, but there are a number of relational magnitudes which might serve, and progress is being made in discovering them.

If the momentary retinal pictures, shifting with every new fixation of the eyes, are taken to be the stimuli for vision, as physiologists assume, the fact of permanence in perception borders on the miraculous. The situation is clarified when it is realized that retinal images are not stimuli but incidents in the exploratory activity of the eyes, which function to register the ambient light at a given location. There is a sequence of overlapping images, to be sure, for the exploration must operate overtime, but the particular sequence is largely irrelevant. The stimulus for the visual system is focusable light, and this is stable and permanent for the given location. The situation is not unlike that of exploratory touch, with the hand and fingers moving over the edges of an object. The changing pressure-patterns on the skin appear to be wholly incoherent and, indeed, the sequence may never be twice the same. But the invariant features of the transforming pressure-patterns are perfectly coherent, and these are registered. These specify the permanent shape of the object.

**Changes and events.** Within the permanent environment described above, there may be moving objects, and their motion may be either rigid or nonrigid. Substances in the solid state generally move without deformation, in accordance with mechanics; liquids and gases (streams, clouds, fire) generally undergo deformation. The motions of animals, including the expressive movements of human beings, are of the latter sort. Both types and many subvarieties of such movements are distinguished by vision. The projected motions which present themselves to an eye are all geometrical transformations. With this as a basis, the question

is how the rigid motions of objects can be distinguished from the elastic changes. Putting the question differently, why are the perspectives of the same object not confused with the changes that would transmute an object into a different one? The answer may be that the transformations in question belong to different "groups," and that the eye registers the difference. In fact, both human and animal eyes seem to be very sensitive to the types and parameters of geometrical transformation.

When a door opens, the optical rectangle becomes a trapezoid but the door is seen as rigid and turning. The variant component of the stimulus yields change of slant; the invariant component yields rigidity or constant shape. When an ameba elongates itself, the same invariants are not present; elasticity or changing shape is seen.

The perception of even more complex events of the natural environment, such as the collision of two solid objects, is beginning to be studied. The optical stimulus for the impression of one-thing-moving-another can be isolated, and as A. Michotte (1954) has demonstrated, the experience is more simple and direct than might be anticipated.

**Persons and the cultural environment.** In addition to the environment of posture, locomotion, and manipulation, there is the complex environment of social behavior. Not only solid objects but social objects are perceived. Other individuals and their actions, including facial expressions, gestures, speech, writing, pictures, music, and symbols, are all sources of visual and auditory stimulation. This is the area in which a great deal of research in perception has been concentrated because these stimuli and the corresponding perceptions preoccupy human life. But this is also where the perceptual process becomes most complicated and least direct. The pattern and sequence of the stimulus ceases to be geometrically related to its object or event and begins to be arbitrarily related to it only by social agreement. The optical forms and acoustical sounds begin to be coded, that is, to have meaning by convention rather than by isomorphic relation. Signals or symbols will induce correct percepts, but only if the perceiver has been educated in the social group which uses them.

The perception of persons may be fairly direct, as are the perceptions of even simple actions like attack, but the perception of their words is indirect. The understanding of language is inseparably related to the producing of language, that is, to communicating. For the child, making speech sounds or written marks is part of perceiving them. Both must be learned, the spoken word along with the perception of speech, and writing along with reading. The child thus learns to evoke perceptions in another individual as well as to have perceptions. Communication is usually circular, and there is a feedback loop through another individual beyond the kinesthetic, auditory, and visual loops of self-stimulation. The ability to respond to the products of one's own vocal or graphic behavior in the same

way as others do is basic to human knowledge and thought. When the child begins to perceive things in cooperation with others, the perceptual process reaches a higher stage than before.

There are many references in the scientific literature of experimental work on the perception of graphic material, that is, line drawings, pictures, symbols and nonsense forms. But this work is hard to interpret because stimulus material carries a mixture of intrinsic meaning and coded meaning. Visual forms or line drawings have a vague status which lies somewhere between substantial objects and symbolic tracings on a surface. Experiments in "form perception" suffer from this ambiguity; the observer tends to perceive both a semirepresented real object and the literal markings on the surface, the percept being a compromise between them. The light entering the eye does not specify either a fully substantial surface or a fully symbolic artifact, but it has some features of both.

Perceiving by way of coded stimuli is a large field for experimental investigation. The most active study up to the present has been connected with communication systems and with verbal messages (see INFORMATION THEORY). The study of partially coded stimuli, facial expressions, or music, for example, is less advanced. And the scientific understanding of the nonrepresentative pictures made by modern artists is at a primitive level.

Perceiving is in part a matter of first-hand acquaintance, independent of language. It is also in part a kind of experience at second hand, modulated by the conventions of language. The blending of these seemingly disparate influences in man's knowledge of the world is a major theoretical problem which is not yet solved. [J.J.G.]

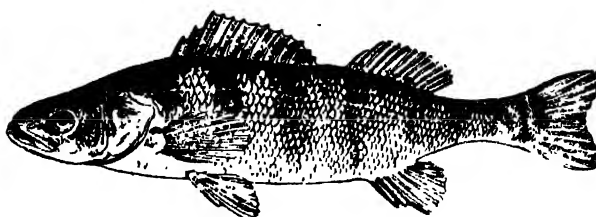
**Bibliography:** E. G. Boring, *Sensation and Perception in the History of Experimental Psychology*, 1942; E. Brunswik, *The Conceptual Framework of Psychology*, 1952; J. J. Gibson, *The Perception of the Visual World*, 1950; J. J. Gibson, J. Putdy, and L. Lawrence, A method of controlling stimulation for the study of space perception: the optical tunnel, *J. Exptl. Psychol.*, 50:1-14, 1955; K. Koffka, *Principles of Gestalt Psychology*, 1935; A. Michotte, *La Perception de la Causalité*, 2d ed., 1954; M. D. Vernon, *A Further Study of Visual Perception*, 1954.

## Perch

A term which, when properly limited, applies only to the yellow perch, *Perca flavescens*, and its European relatives.

The yellow perch is one of the most common of the panfishes in North America, especially in the Upper Mississippi and Great Lakes drainage areas. It is a fine food fish, readily caught by anglers at all seasons of the year. In many lakes it is given to over-reproduction and stunting if not severely fished.

The term perch is commonly misused. For example, the rosefish is frequently called red perch or



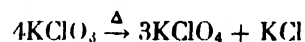
The yellow perch, *Perca flavescens*; length to 12 in. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

ocean perch; the white bass or fresh-water drum are often called white perch; and the general term perch is used for any of several fresh-water sunfishes, particularly black perch or the green sunfish. See PERCIFORMES. [J.D.B.]

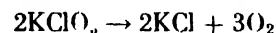
## Perchlorate

A compound which contains chlorine in the 7+ oxidation state and which is derived from perchloric acid,  $\text{HClO}_4$ . Perchlorates are more stable than chlorates, chlorites, or hypochlorites but are nevertheless excellent oxidizing agents. On heating, perchlorates decompose into potassium chloride,  $\text{KCl}$ , and oxygen gas. Because of their oxidizing properties, perchlorates find use in explosives and as oxidizing agents in the laboratory.

Potassium perchlorates can be prepared by heating potassium chlorate.



In a side reaction some oxygen is liberated.



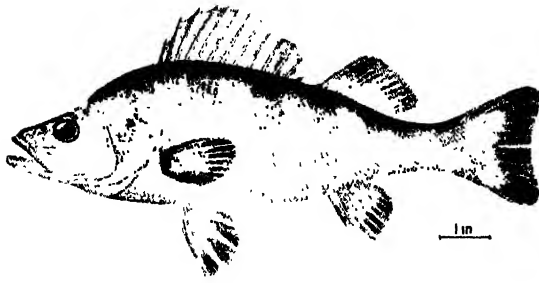
Sodium perchlorate is also prepared commercially by an electrochemical process. Sodium chlorate made from sodium chloride and chlorine is the starting material (see CHLORATE). The sodium chlorate is then electrolyzed with the net reaction at the anode being



Sodium perchlorate is then converted to perchloric acid or other metallic salts. See CHLORINE; HYPOCHLORITE; OXIDIZING AGENT. [E.E.WR.]

## Perciformes

The typical spiny-rayed fishes, also known by the ordinal names Acanthopteri and Percomorphi. This is the largest order of vertebrates. It includes a diversity of structural types as well as of size, from a length of less than  $\frac{1}{2}$  in. to a weight of nearly 1 ton. The characters of the Perciformes include fin spines which are usually present; a pelvic fin which, if present, is usually thoracic or jugular in position; the pelvic girdle usually attached to the cleithra, sometimes connected by ligaments; the pelvic fin usually with a spine and 5 soft rays, the latter occasionally reduced; the pectoral fin base more or less vertical, usually placed well up



Yellow perch, *Perca flavescens*. (After G. B. Goode, *Great International Fisheries Exhibition, London, 1883*, U.S. Natl. Museum Bull. 27)

on the side; an upper jaw bordered largely or entirely by premaxillae; orbitosphenoid and meso-coracoid absent; a swimbladder without a duct; a posttemporal which is usually forked, articulating to the skull; scales usually ctenoid, sometimes secondarily cycloid, absent, or variously modified; caudal fin with 17 principal rays (15 branched) or fewer; hyoid arch with 4 branchiostegal rays attached to the outer face of the epihyal and ceratohyal above the prominent angle of the ceratohyal, plus 1-4 rays, usually 2-3, attached to the edge of the ceratohyal below the angle, the number rarely further reduced.

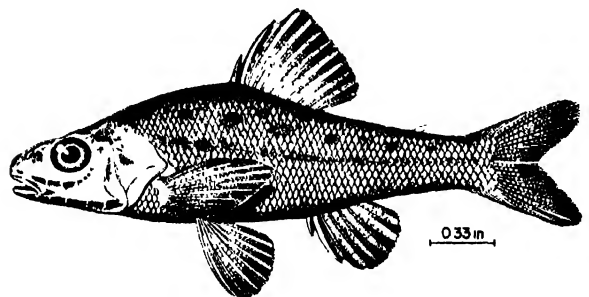
**Adaptive radiation.** Perciform fishes dominate the modern vertebrate life of the oceans and have done so throughout the Cenozoic. The group first appeared in the Upper Cretaceous, after which it underwent a rapid adaptive radiation; many of the basic structural types, as well as most major orders of perciform derivatives such as the Pleuronectiformes, Tetraodontiformes, and Lophiiformes, were present in the Eocene. A few families of perciforms have been notably successful in fresh water, the Cichlidae in Africa and South America, the Centrarchidae in North America, the Percidae in North America and Eurasia, the Anabantidae in southeastern Asia, and many other families which have achieved limited success in invading fresh waters. Other families including the Chiasmodontidae, Brotulidae, and Cyclopteridae have effectively adapted to life in the deep seas, and still others, such as the Scombridae, Stromateidae, and Coryphaenidae, have become specialized for pelagic existence. It is in the shore areas, the offshore banks, the coral reefs, the coastal beaches and lagoons, and the intertidal zone, however, that the perciforms have attained their ultimate achievement. Here the enormous variety attests the adaptive effectiveness of the group. Because the knowledge of classification is still imperfect, no accurate enumeration is yet possible, but a rough tally indicates that the order contains 17 suborders, about 137 families, and nearly 1200 genera. These figures are conservative; the number would be substantially increased by adopting the arrangement accepted by splitters. An estimate of the number of included species, perhaps 8000, should be accepted merely

as a rough guess, possibly in error by 25% or even more.

**Economics.** From an economic standpoint the Scombridae, including the oceanic tunas and mackerels, rank first among the perciforms. The Sciaenidae or drums, the Serranidae or sea basses, the Scorpaenidae or rockfishes, the Carangidae or jacks, the Cichlidae of tropical fresh waters, the Percidae of temperate fresh waters, and other groups support important commercial fisheries. Some of these and other families, such as the Centrarchidae or sunfishes, the Istiophoridae or sailfishes and marlins, and the Pomatomidae or bluefishes are valued in sport fisheries, and hence are of great recreational and indirect economic importance. See ACTINOPTERYGII; FISHERIES CONSERVATION. [R.M.B.]

## Percopsiformes

A small order of actinopterygian fishes which is perhaps remotely related to the Beryciformes. The group is also known as the Salmopercae. Its characters include single ray-supported dorsal and anal fins, each with 1-4 anterior spines; pelvic fin subabdominal in position, with a minute spine and 7-8 soft rays; pelvic girdle attached to the postcleithra; upper jaw bordered by premaxillae; no orbitosphenoid bone; swimbladder without a duct; and body covered with ctenoid scales.



Sand roller, *Percopsis transmontana*. (After D. S. Jordan and B. W. Evermann, *The Fishes of North and Middle America*, U.S. Natl. Museum Bull. 47, 1900)

Two families, two or three genera, and three Recent species of North American fresh-water fishes comprise the order. Eocene and Miocene fossil genera from North America are assigned to the group. The species attain a maximum length of 6 in. and inhabit sluggish or standing waters. See ACTINOPTERYGII; BERYCIFORMES. [R.M.B.]

## Perennial plants

Plants which live for more than two years. Trees and shrubs, and many grasses, sedges, and other herbaceous plants are perennials. During unfavorable seasons the aerial parts may die, but the roots remain alive. Some perennials produce flowers and seeds during their first year and annually thereafter. Some, such as the apple, bear no flowers until their fourth or fifth year but bloom each year thereafter. Other perennials, as the agave, bear no flow-



ers before they become 10 or more years old, and then, after flowering, the plants usually die. See PLANT. [P.D.S.]

## Pericycle

The pericycle is commonly defined as the outer boundary of the stele of plants (see STELE). Originally it was interpreted as a band of cells between the phloem and the innermost layer (endodermis) of the cortex. Such pericycle is commonly found in roots and, in lower vascular plants, also in stems. In higher vascular plants, however, a distinct layer of cells may not be present between the phloem and the cortex. The homogeneous groups of fibers located on the outer boundary of the primary phloem in many stems are not pericyclic in origin but develop in the earliest part of the phloem, whose remaining cells are obliterated. The pericycle, if present, may be composed of parenchyma or sclerenchyma cells with relatively thin or heavily thickened walls. It may be one to several layers in radial dimensions (Fig. 1).

Primordia of branch roots commonly arise in the pericycle in seed plants, most frequently outside the xylem ridges (Fig. 2). The first cork cambium may also arise in the pericycle of those roots that

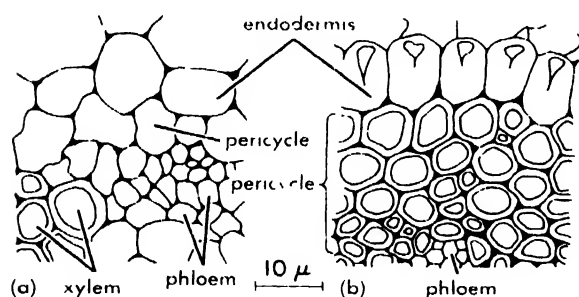


Fig. 1. (a) Part of transection of root of *Actaea alba* Mill., including xylem and phloem. Pericycle thin-walled and one cell in radial dimension. (b) Part of transection of root of *Smilax herbacea* L. including phloem. Pericycle thick-walled and 4–5 cells in radial dimension (bracket).

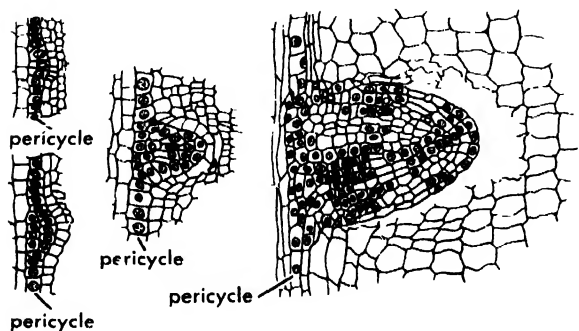


Fig. 2. Vertical section of a root showing progressive stages in development of a secondary root from the pericycle, and growth through the cortex of the primary root. (From G. M. Smith, *A Textbook of General Botany*, 5th ed., Macmillan, 1953)

have secondary vascular tissues. In roots, a part of the vascular cambium itself (that outside the primary xylem ridges) originates from pericycle cells. See CORTEX, PLANT; ENDODERMIS; MERISTEM, LATERAL; PARENCHYMA; PHLOEM; ROOT (BOTANY); SCLERENCHYMA; STEM (BOTANY); XYLEM; see also PLANT TISSUE SYSTEMS. [V.I.C.]

## Periderm

The protective tissue system of stems and roots composed of the cork cambium (phellogen) and its derivatives cork (phellem) and phelloderm (parenchyma). Cork occurs on the outside, the phelloderm on the inside of the cork cambium (Fig. 1). Periderm develops typically on woody roots and stems of dicotyledons and gymnosperms. Herbaceous stems, stems of monocotyledons, leaves, and even fruits may likewise develop periderm although it may be limited in area and thickness. Some woody monocotyledons develop a special kind of periderm, corklike in structure, called storied cork because of cell arrangement. Corky zones of many herbaceous roots are really suberized cortex.

**Origin of periderm.** Typical periderm originates in more or less mature primary tissues, the cells of which resume meristematic activity. The divisions occur parallel with the surface of the organ, that is, in tangential planes. In roots, the cork cambium usually arises in the pericycle and the subsequent formation of cork causes the sloughing of the cortex. In stems, any living tissue outside the vascular cambium may develop a cork cambium. Cells produced by this cambium toward the outside differentiate into cork as the cell walls become suberized and the protoplasts disintegrate. Some cork cells appear to be empty, others contain resinous or tanniferous ergastic substances. The phelloderm cells, which are formed toward the inside, remain alive and often resemble the parenchyma cells of cortex or phloem. Their radial alignment with the cork cambium and cork provides the evidence that they are part of the periderm. The phelloderm is sometimes called secondary cortex.

Lenticels are characteristic of many periderms (Fig. 2). These structures are lens-shaped regions of periderm where the cork cambium divides more frequently than elsewhere and the derivative cells are not as compactly arranged as the cork cells. The lenticels may develop before the rest of the periderm; the cork cambium then spreads from lenticel regions. The lenticel cells may or may not be suberized. All or part of these cells gradually separate from one another. In many plants layers of loosely arranged cells alternate with compact layers, which are ruptured periodically by the production of more cells underneath. The cells in the alternating zones are called complementary cells and closing cells, respectively. Functionally, lenticels are considered to be regions of the periderm that allow gaseous interchange between the internal regions of the plant and the atmosphere. In this respect, lenticels seem to be secondary replacements of the stomata of the epidermis.

**Cork.** Cork may be elastic, smooth, and uniform in structure. Or it may also contain sclereids (hard-walled cells) distributed in various patterns. Cork varies in thickness. Commercial cork is an example of a thick cork. Annual rings are known to occur in

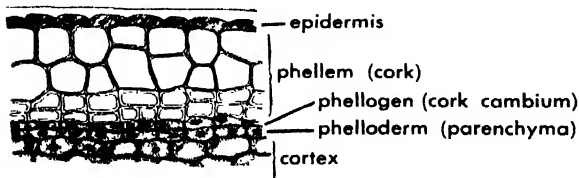


Fig. 1. Periderm. Cross section of superficial layer in twig of *Populus deltoides*; the cork cambium arose in the outermost cortical cells, and has formed four layers of cork cells and one of phelloderm; the cork is covered by dead, tannin-filled epidermal cells. (From A. J. Eames and L. H. MacDaniels, *An Introduction to Plant Anatomy*, 2d ed., McGraw-Hill, 1947)

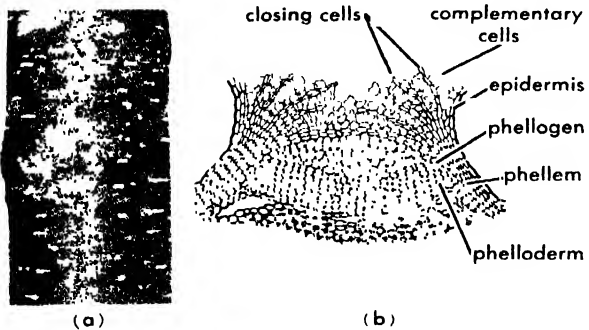


Fig. 2. (a) Lenticels of alder (*Alnus incana*). The stem was 4 in. in diameter (from C. L. Wilson and W. E. Loomis, *Botany*, rev. ed., Dryden, 1957). (b) Lenticel of *Prunus avium* in transverse section of stem. A number of successive layers of complementary and closing tissue have been formed, and the thick layer of phelloderm dips inward into the cortex (from A. J. Eames and L. H. MacDaniels, *An Introduction to Plant Anatomy*, 2d ed., McGraw-Hill, 1947).

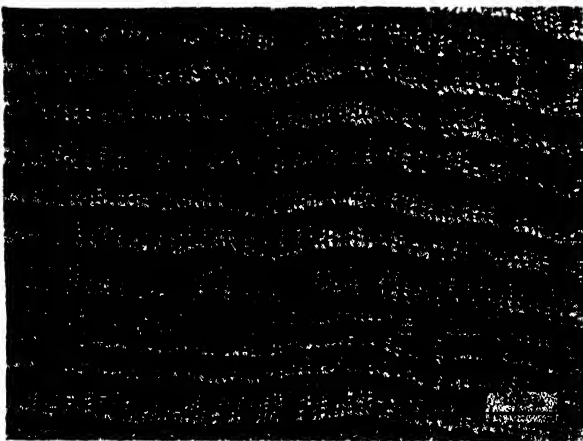


Fig. 3. Radial section of periderm of paper birch (*Betula papyrifera* Marsh) showing alternate layers of radially narrow and broad phellem (cork) cells. (Forest Products Laboratory, USDA)

cork (Fig. 3). Some plants produce corky wings on their stems, for example, cork elm and sweetgum. Roots and underground stems rarely accumulate cork. Exposed portions of roots or rhizomes, on the other hand, may accumulate cork (or rhytidome) in much the same way as aerial stems.

Corky zones develop regularly as spots on some fruits, for example, apple and pear, in the styles of pineapple, and as ridges on *Circaea* fruits. In some aquatic or marsh plants, cork cells become T-shaped, so that the tissue encloses large air spaces. Such tissue is an example of the so-called aerenchyma. Cork which forms in response to wounding is similar to normal cork. Periderm often develops in connection with abscission of leaves and other structures; it may precede this phenomenon or it may develop as a wound cover after abscission. See BARK; CORTEX, PLANT; EPIDERMIS, PLANT; FRUIT (BOTANY); LEAF (BOTANY); PARENCHYMA; PERICYCLE; PHLOEM; ROOT (BOTANY); SCLERENCHYMA; STEM (BOTANY). [H.W.BL.]

## Peridotite

A dark-colored, phaneritic (visibly crystalline) igneous rock composed largely of olivine with smaller amounts of pyroxene or hornblende. It is one of a group of essentially feldspar-free rocks called ultramafites (meaning extremely rich in mafic or dark-colored minerals). As feldspar (calcic plagioclase) increases, the rock passes into olivine gabbro; as pyroxene increases at the expense of olivine, the rock passes into pyroxenite. As hornblende becomes the dominant mafic, the rock becomes a hornblendite. Fresh peridotite is one of the heaviest (specific gravity  $3.3 \pm$ ) igneous rocks.

**Composition.** Olivine is generally magnesium-rich and forms rounded to irregular grains. It is essentially the only mineral in the green, sugary-looking rock called dunite. Orthopyroxene is generally magnesium-rich (enstatite), and the clinopyroxene may be diopside, augite, or rarely titaniferous augite. Amphibole is brown or green hornblende and less commonly barkevikite. Mica is magnesium-rich (phlogopite) and colorless to red-brown. Accessory minerals include chromite, magnetite, ilmenite, spinel, apatite, and garnet. A variety of mica peridotite called kimberlite occurs in pipelike bodies near Kimberly, South Africa, and carries small quantities of diamond. Other peridotites are platinum-bearing. Peridotite is usually more or less altered to serpentine, chlorite, carbonate, talc, and actinolite.

**Texture and structure.** Most peridotites are even-grained except a few which show large, irregular pyroxene grains enclosing smaller grains of olivine and magnetite. Certain minerals may be segregated to form banded and layered structures. In some rocks elongate olivine grains in subparallel orientation produce a flow structure. In others the texture suggests extensive movement and granulation of olivine crystals.

**Occurrence.** Peridotites are found as thin sheets or lenses interlayered with gabbro, pyroxenite, and anorthosite. Here they appear to have differenti-

ated from gabbroic magma (rock melt). Fine examples include those of the Stillwater complex in Montana and the Transvaal, South Africa. Where formed as pipes, plugs, and dikes, peridotite appears to have had little thermal effect upon the enclosing rocks. This suggests the material was injected as a relatively cool crystal mush rather than a complete melt (magma). Many serpentine masses in fold-mountain chains may represent altered peridotite intrusives. Perhaps many such bodies, however, represent products of metamorphism and metasomatism. See GABBRO; IGNEOUS ROCKS; METAMORPHISM; METASOMATISM; PETROGRAPHIC PROVINCE; PYROXENITE. [C.A.C.A.]

### Perigee

The point nearest the Earth in the orbit of the Moon or of an artificial satellite. At perigee the Moon is  $5\frac{1}{2}\%$  closer to Earth than at its mean distance, the orbital eccentricity being 0.055. Because, on the average, the Moon and Sun subtend nearly equal angles, a solar eclipse near perigee lasts about 5 min; an eclipse near apogee is annular. The line perigee-Earth-apogee is the major axis of the orbital ellipse or the line of apsides. The differential attraction of the Sun on Earth and Moon causes the line of apsides of the Moon to move forward in the orbital plane with a period of 8.85 years. See MOON; PERIHELION. [C.P.K.]

### Perihelion

In astronomy, that point at one extremity of the major axis of the elliptical, parabolic, or hyperbolic orbit of a planet or comet about the Sun, where the planet or comet is closest to the Sun. The instant when a planet or comet is at perihelion is referred to as the time of perihelion passage. For Earth this occurs about the third of January, at which time Earth is some 1,550,000 miles closer to the Sun than its mean distance of 92,900,000 miles. See CELESTIAL MECHANICS; ORBITAL MOTION. [R.L.D.]

### Period (periodic phenomena)

The time interval of a single repetition of a varying quantity of a motion or phenomenon which repeats itself regularly. The period is the reciprocal of the frequency. See FREQUENCY (WAVE MOTION).

Waves which have regularly repeated time-varying quantities are termed periodic (see PERIODIC MOTION). In general, any complex periodic wave can be described by a Fourier analysis as the sum of a series of sine and cosine partial waves whose periods are integral multiples of a single period known as the fundamental period of the complex wave. See WAVE MOTION. [W.J.C.]

### Periodate

A salt which contains iodine in the 7+ oxidation state and which is derived from periodic acid. Periodic acid is known in three forms: metaperiodic acid,  $\text{HIO}_3$ ; dimesoperiodic acid,  $\text{H}_4\text{I}_2\text{O}_9$ ; and paraperiodic acid,  $\text{H}_5\text{IO}_6$ . The corresponding salts are also known.

Only the periodates of sodium and potassium are important. Sodium metaperiodate,  $\text{NaIO}_4$ , and potassium metaperiodate,  $\text{KIO}_4$ , are used as oxidizing agents in analytical chemistry. See IODINE.

[E.E.WR.]

### Periodic motion

Any motion that repeats itself identically at regular intervals. If  $x(t)$  represents the displacement of any coordinate of the system at time  $t$ , a periodic motion has the property that

$$x(t + T) = x(t)$$

for every value of the variable time  $t$ . The fixed time interval  $T$  between repetitions, or the duration of a cycle, is known as the period of the motion. Frequency is the number of repeating cycles per unit time, and is numerically equal to the reciprocal of the period  $T$ .

The motion of the escapement mechanism of a watch, the motion of the earth about the sun, and the more complicated motion of the crankshaft, piston rods, and pistons in an engine running at uniform speed are all examples of periodic motion.

The vibration of a piano string after it is struck is a damped periodic motion, not strictly periodic according to the definition. Although the motion very nearly repeats itself, and with a fixed repetition time, each successive cycle has a slightly smaller amplitude. See DAMPING.

Any periodic motion can be expressed as a Fourier series—a sum of sine and cosine terms whose frequencies are integral multiples of the frequency  $f$  of the periodic motion. Thus

$$x(t) = A_0 + \sum A_n \cos(2\pi nft) + \sum B_n \sin(2\pi nft)$$

where the  $A$ s and  $B$ s are constant coefficients, and the sums may be taken over all positive integer values of  $n$ . For the special case in which the coefficients all vanish for  $n > 1$ , see HARMONIC MOTION; see also FOURIER SERIES AND INTEGRALS.

Many systems with more than one degree of freedom, whose motion is not simply periodic, are multiply periodic. The motion may be resolved into parts (for example, horizontal and vertical components, radial and tangential components) each of which is periodic, but with periods that are not commensurate. One example is the vibration of a bell, whose overtone frequencies are not simply related to the fundamental frequency. The motion of the solar system is multiply periodic because it never exactly repeats itself, even though each planet moves periodically. See VIBRATION; WAVE MOTION. [J.M.KE.]

### Periodic table

A table of the elements, written in sequence in the order of atomic number or atomic weight and arranged in horizontal rows (periods) and vertical columns (groups) to illustrate the occurrence of similarities in the properties of the elements as a periodic function of the sequence.

Although the principle of the periodic arrangement was firmly established nearly a century ago,

PERIODIC TABLE OF THE ELEMENTS																VII a					
I a												II a		III a		IV a	V a	VI a			
1 H 1.0080														5 B 10.82	6 C 12.011	7 N 14.008	8 O 16.000	9 F 19.00	10 Ne 20.183		
3 Li 6.940	4 Be 9.013													13 Al 26.98	14 Si 28.09	15 P 30.975	16 S 32.066	17 Cl 35.457	18 Ar 39.944		
11 Na 22.991	12 Mg 24.32	III b	IV b	V b	VI b	VII b	VIII		I b		II b										
19 K 39.100	20 Ca 40.08	21 Sc 44.96	22 Ti 47.90	23 V 50.95	24 Cr 52.01	25 Mn 54.94	26 Fe 55.85	27 Co 58.94	28 Ni 58.71	29 Cu 63.54	30 Zn 65.38	31 Ga 69.72	32 Ge 72.60	33 As 74.92	34 Se 78.96	35 Br 79.916	36 Kr 83.80				
37 Rb 85.48	38 Sr 87.63	39 Y 88.91	40 Zr 91.22	41 Nb 92.91	42 Mo 95.95	43 Tc (99)*	44 Ru 101.1	45 Rh 102.91	46 Pd 106.4	47 Ag 107.868	48 Cd 112.41	49 In 114.82	50 Sn 118.70	51 Sb 121.76	52 Te 127.61	53 I 126.91	54 Xe 131.30				
55 Cs 132.91	56 Ba 137.36	57 La 138.92	72 Hf 178.50	73 Ta 180.95	74 W 183.86	75 Re 186.22	76 Os 190.2	77 Ir 192.2	78 Pt 195.09	79 Au 197.0	80 Hg 200.61	81 Tl 204.39	82 Pb 207.21	83 Bi 208.99	84 Po 210.	85 At (210)*	86 Rn 222.				
87 Fr (223)*	88 Ra 226.05	89 Ac 227.0																			
LANTHANUM SERIES			58 Ce 140.13	59 Pr 140.91	60 Nd 144.27	61 Pm (147)*	62 Sm 150.35	63 Eu 152.0	64 Gd 157.26	65 Tb 158.93	66 Dy 162.51	67 Ho 164.94	68 Er 167.27	69 Tm 168.94	70 Yb 173.04	71 Lu 174.99					
ACTINIUM SERIES			90 Th 232.05	91 Pa 231.	92 U 238.07	93 Np (237)*	94 Pu (242)*	95 Am (243)*	96 Cm 247.*	97 Bk (249)*	98 Cf 251.*	99 Es 254.*	100 Fm (253)*	101 Md (256)*	102 No (253)*						

\*mass number of most stable known isotope

there has been much debate as to the most suitable form of display. A widely accepted modern arrangement is presented in the accompanying table, which is commonly referred to as a "long" table.

Each element, represented by its symbol, occupies a separate square of the table, together with its atomic number and atomic weight. The sequential arrangement is in the order of atomic number.

The table divides the elements into nine groups, designated by numerical column headings, and seven periods. Seven of the nine groups are further divided into a and b categories, the a elements often being classified as main group and the b as subgroup elements. Two rows of elements (the lanthanide, or rare-earth, series and the actinide series) which on the whole are best classified as members of group IIIa occupy special positions outside the main body of the table, as they cannot be included conveniently in periods six and seven.

The following sections serve to illustrate some of the correlative features of periodicity.

**Valence.** In general, elements in the same group display a similar valence, which is numerically equal to the group number. The rule holds most firmly for the main groups I through V, less so for the subgroup elements, and still less for the elements of groups VI, VII, and VIII. Among these latter, the group number valence is more readily achieved by the heavier, as compared with the lighter, elements. (Group 0 elements are distinguished by an almost complete lack of chemical reactivity and hence show zero valence.)

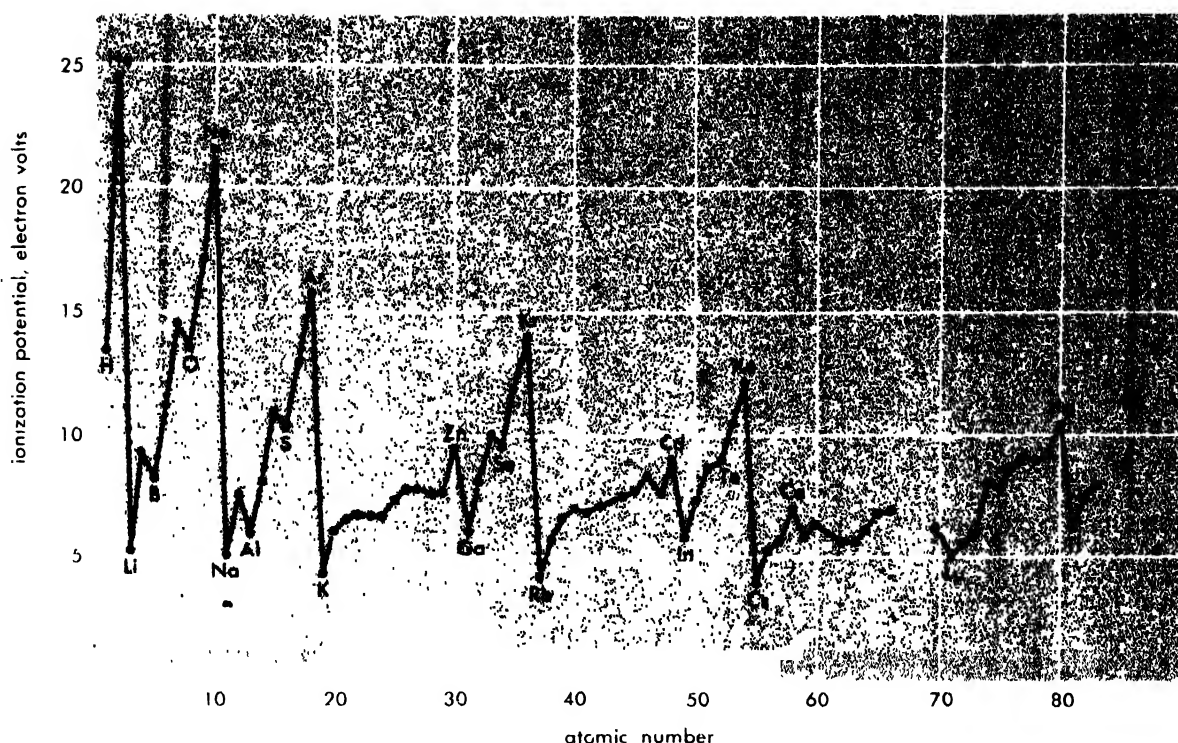
**Metals and nonmetals.** The periodic table effects a natural division of the elements in their elemental or uncombined state into metals and nonmetals. The nonmetals are confined to the lighter elements of groups IVa through VIIa. Between the area occupied by the typically metallic elements

and that occupied by the nonmetals there is a somewhat ill-defined borderland of elements (germanium Ge, arsenic As, antimony Sb, tellurium Te, and polonium Po) whose properties are transitional between the metallic and nonmetallic elements.

**Acid-base properties of the oxides of the elements.** Associated in a general way with the degree of metallic character of an element is a property called electropositivity, which expresses the tendency of an element to form positive ions by losing electrons. This property is exhibited in highest degree by the heaviest elements of group I and diminishes generally on proceeding to elements which lie above and to the right of this position. Highly electropositive elements form oxides or hydroxides which are strong bases. Potassium hydroxide (caustic potash) is a familiar example. The nonmetals, such as sulfur, attract electrons and hence are termed electronegative. Solutions of the oxides of these elements in water yield acids. Sulfur trioxide, for example, reacts with water to form sulfuric acid.

**Atomic volumes.** The volume occupied by 1 gram-atomic weight of an element in the solid state is called its atomic volume. If this property is plotted against atomic number, the resultant curve exhibits periodic maxima and minima, the former being due principally to the metals of group I and the latter to the metals of group VIII.

**Other properties.** The preceding sections afford only a limited illustration of the immense correlative power of the periodic arrangement. Other properties which have been shown to exhibit analogous trends and periodicities include oxidation potential, heat of formation of type compounds, electrical conductivity, melting point, boiling point, ionic radius, ionization potential, electron affinity, optical spectrum, and magnetic behavior.



First ionization potentials of the elements plotted as functions of atomic number.

Although some mention has been made of the lack of agreement with respect to details of the periodic arrangement, it must not be imagined that its broad systemizations are more dependent upon one form of table than another.

The periodic table was developed and largely perfected by Dmitri Mendeleev in the mid-nineteenth century as an empirical correlation between the chemical properties of the elements and their atomic weights. In the light of modern knowledge, no such correlation is to be expected, since the atomic weight is determined almost entirely by the mass of the atomic nucleus which plays only an insignificant role in chemical reactions.

With few exceptions, however, the sequence of elements according to atomic weight is identical with that according to atomic number. This latter quantity is numerically equal to the number of extranuclear electrons in the neutral, or uncharged, atom of an element. It is the electrons, particularly those furthest removed from the nucleus, which determine chemical behavior. The laws of atomic architecture, discovered in the early part of the present century, require that in a sequence of atoms of regularly increasing atomic number, there should be a periodic recurrence in the number and type of electrons in the outermost electronic shell. This forms the true basis of the periodic variation of the properties of the elements.

Few systemizations in the history of science can rival the periodic concept as a broad revelation of the order of the physical world. In the rhythmic pattern of the properties of the elements, the architectural units of the universe, no aspect of

their behavior changes in a capricious or wholly novel way. Whatever new elements may be discovered in the future, it is certain they will find a place in the periodic system, conforming to its order and exhibiting the proper familial characteristics.

See ATOMIC STRUCTURE AND SPECTRA; VALENCE.  
[B.B.C.U.]

*Bibliography:* J. H. Hildebrand and R. E. Powell, *Principles of Chemistry*, 6th rev. ed., 1952; T. Moeller, *Inorganic Chemistry*, 1952; L. Pauling, *College Chemistry*, 2d ed., 1955; G. T. Seaborg and E. G. Valens, *Elements of the Universe*, 1958.

## Periodicity in organisms

Characteristic rhythms or cycles of living phenomena. Collectively, living organisms display a broad spectrum of period lengths ranging from such high frequencies as those of brain waves (a complex of electrical-potential rhythms in the brain with many cycles per second) and heart beat, to the several-year cycles of abundance of various species. However, periodicity in organisms shall here be arbitrarily interpreted to include only a limited group of periods, those related to the natural physical ones of the earth, namely days, lunar tides, months, and years. These periodicities commonly comprise important adaptations of the organisms to the normal fluctuations of such obvious environmental factors as illumination, temperature, and the ocean tides. These rhythms, which in organisms in nature are usually quite precisely of the natural geophysical frequencies, have certain peculiar properties in common which tend to set them apart from other

biological periodisms and support the view that they are utilized very importantly as "clocks" and "calendars," to enable living things to live more harmoniously with their rhythmic physical environment. *See* **ECOLOGY**.

Examples of solar-day (24-hour) rhythmic phenomena, related to the rotation of the earth relative to the sun, are the sleep movement of plant leaves and petals, change in skin color of crabs, emergence of fruitflies from pupal cases, wakefulness and spontaneous activity of animals, body temperature, hormone titers in the blood, and susceptibility to drugs and toxic agents. Lunar-day (24.8-hour) rhythms, related to the rotation of the earth relative to the moon, are characteristic of many activities of animals dwelling on the seashores, subjected continuously to the moon-dominated ocean tides. The rhythmic daily and tidal patterns of organismic change are adjusted to the physical environmental cycles so as to be of optimal survival value. The synodic month of 29.5 days is the period between two successive synchronizations of solar and lunar "noons," or the average period separating two consecutive new moons. Of this average period are, for example, the reproductive cycles of many marine animals and plants, and the human menstrual periods. Annual periodicities in organisms are most evident at latitudes distant from the Equator, but strangely enough, they may also occur in organisms in the annual constancy of equatorial regions. Familiar examples of annual rhythms are those of reproduction in animals and of growth, flowering, and fruiting in plants. *See* **BODY RHYTHM**; **ESTRUS**; **GESTATION PERIOD**; **PLANT PHYSIOLOGY**.

**Persistence of the rhythms.** Although in the natural environment the observed cycles clearly depend in good measure upon organismic responses to the cyclic patterns of such factors as light, temperature, and the ocean tides, it is commonly found that when the organisms are removed from their habitat and placed in conditions constant with respect to these factors, the organism not only may continue to exhibit the same rhythmic fluctuations with closely the same periods, but may retain in the daily or tidal cycles even detailed characteristics of the environmentally impressed adaptive pattern. The organism therefore possesses some means for timing these cycle periods which does not depend upon the rhythms of the obvious environmental factors which normally determine the form of pattern of the cyclic fluctuation. Furthermore, this whole pattern of the persistently recurring cycles can be abruptly shifted to bear any arbitrarily selected phase relationship to the natural external day-night cycles by appropriately adjusting the time of day of occurrence of changes in such factors as light or temperature in artificially administered 24-hour cycles of these factors. Once shifted, the cycles may persist in the constant laboratory conditions with the newly impressed relationships relative to the time of day by the clock. It is also known that even in constant conditions of

light and temperature, the phases of the recurring fluctuations may display a spontaneous small daily shift to yield observed regular periods which are longer or shorter than the natural ones.

The rhythms persisting in constant conditions are timed by a mechanism well adapted for retaining temporal precision. If timing were dependent exclusively on a clock of a conventional metabolic type, the periods would shorten at higher temperatures and lengthen at lower ones. Drugs modifying metabolic rate would similarly give timing inaccuracies. However, numerous experiments have shown the organismic timer to be essentially independent of such alterations in metabolic rate. Color-change cycles with accurate solar and lunar periods continue over a 20°C temperature range in fiddler crabs in constant conditions. Plants continue to show daily sleep movements of their leaves, or other rhythmic activities, with relatively little change in period, although they may be subjected to wide ranges of constant temperatures or following application of metabolic depressants. Seeds exhibit an annual rhythm of germinating capacity when stored in constant conditions, whether at -22°C or +45°C, or even in the continuous presence of toxic agents or absence of oxygen. *See* **CHROMATOPHORE**; **PROTECTIVE COLORATION**.

There is substantial evidence that certain persistent solar-day and lunar day rhythms comprise timing systems enabling animals such as birds and arthropods, which navigate using sun or moon, to correct continuously for the earth's rotation, in order to maintain straight compass directions. There are also reasons to believe that solar-day rhythms serve as timers underlying the well-known organismic response to seasonal changes in day length or photoperiod. *See* **PHOTOPERIODISM IN PLANTS**.

**Mechanism of persistent rhythms.** From the preceding it is evident, on the one hand, that the "clocks" underlying organismic periodicities are remarkably stable and accurate in period length and yet, on the other hand, that the cyclic patterns of change are quite labile and plastic as the organisms adaptively use these basic clocks to adjust themselves readily to changing demands of the environment. In brief, these rhythms comprise one of the most remarkable illustrations of organismic adaptation to their physical environment.

Two basic hypotheses exist for the nature of the organismic timing system. One of these is that the organism possesses, quite independently of its physical environment, natural periods of biochemical oscillation matching closely all the natural geophysical frequencies. The periods are believed to have reached their present lengths with the aid of natural selection and now to be passed on genetically. The major difficulty for this hypothesis is the problem of conceiving an organismic oscillator with such long periods as the geophysical ones, and of the extraordinary stability of period. Also, proof for the existence of such a fully autonomous clock system must await demonstration of persistence of these same periods in organisms in space and away



from all possibly effective geophysical periodisms.

The other hypothesis is that the organismic periods depend upon a continuing response of the organism to pervasive geophysical rhythms. It has now been fully established that periodisms of all the natural geophysical frequencies actually do occur in living things as a response to their physical environment, even under conditions constant with respect to all the factors formerly considered able to influence them. It has also been possible to account quite rationally for all the unusual observed properties of physiological rhythms in terms of such exogenous timing, together with known properties of adaptive phasing of the rhythms by light and temperature. This hypothesis also accounts for the extraordinary temperature- and drug-independence of the rhythm periods. [F.A.B.]

**Bibliography:** F. A. Brown, Jr., The rhythmic nature of animals and plants, *Am. Scientist*, 47(2): 147-168, 1959; J. E. Harker, Diurnal rhythms in the animal kingdom, *Biol. Revs.*, 33:1 52, 1958; H. M. Webb and F. A. Brown, Jr., Timing long-cycle physiological rhythms, *Physiol. Revs.*, 39(1): 127-161, 1959.

### Perischoechinoidea

A subclass of Echinoidea lacking stability in the number of columns of plates comprising the ambulacra and interambulacra. The ambulacral columns vary from 2 to 20, the interambulacral from 1 to 14. Of the four included orders, three are exclusively Paleozoic; the other, Cidaroida, includes both Paleozoic and extant members and is probably ancestral to all other surviving echinoids. See BOTHRIOCIDAROIDA; CIDAROIDA; ECHINOYSTI-TOIDA; ECHINOIDEA; PALAEOINGIDA. [H.B.F.]

### Periscope

An optical instrument that permits viewing along a displaced or deflected axis, providing an observer with the view from a position which may be inaccessible or dangerous. Periscopes range in com-

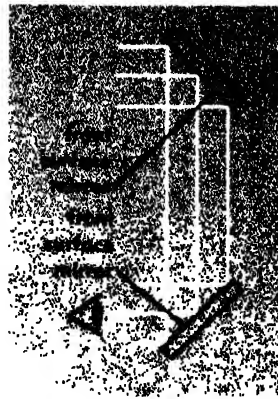


Fig. 3. Simple periscope with mirrors at right angles. Observer views an inverted image.

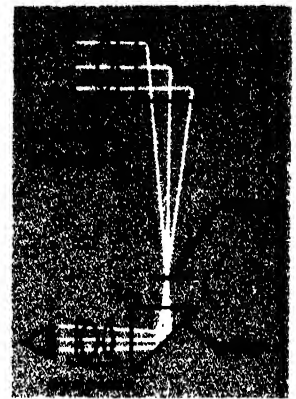


Fig. 4. Rear-sighting periscope with inverting telescope. The image is reinverted.

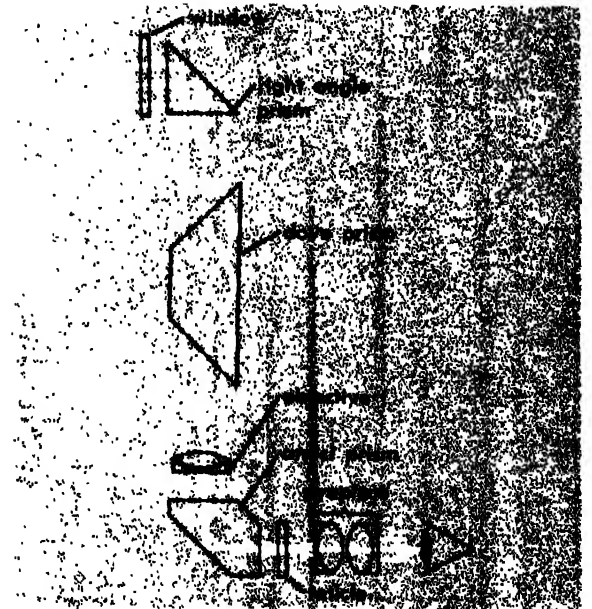


Fig. 5. Panoramic sight.

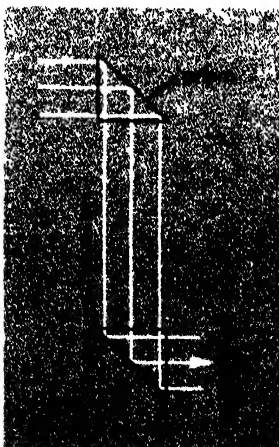


Fig. 1. Simple tank periscope with parallel reflecting surfaces.

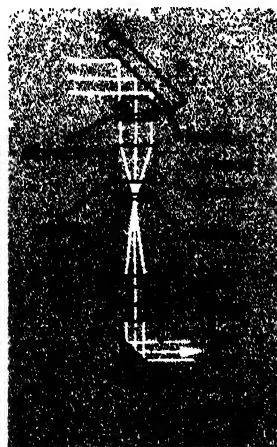


Fig. 2. Tank periscope with terrestrial telescope.

plexity from the simple unit-power tank periscope to the complex multielement submarine periscope.

**Tank periscope.** This device, intended to protect the user from bullets, employs a pair of plane, parallel, reflecting surfaces (either mirrors or prisms), so arranged in a mount that the path of light through the instrument forms a crude letter Z (Fig. 1). If powers greater than unity are desired or if the periscope is to be used for sighting, a terrestrial telescope can be added to the periscope, either as a simple, internally contained system (Fig. 2) or entirely in front of or behind the periscope itself, as desired. The reflecting elements of the system which are responsible for deflecting the optical axis are independent of the refracting (telescope) elements which provide the optical power. See TELESCOPE.

It is also possible to arrange the two mirrors of a periscope at right angles to each other (Fig. 3), in



Fig. 6. Periscopic relay train, showing lenses  $L$ , inversions  $I$ , and angle of view  $\theta$ .

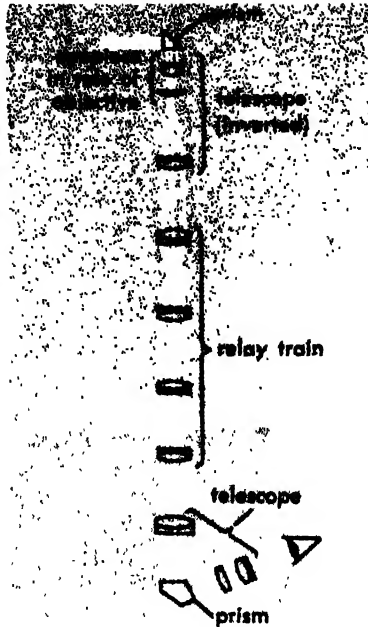


Fig. 7. Submarine periscope, a pair of facing telescopes with a relay train between.

which case the observer views an inverted image with his back to the direction from which light enters the instrument. By adding an inverting (astronomical type) telescope to this system (Fig. 4), the image is reinverted. See TELESCOPE, ASTRONOMICAL.

Periscopes of this type cannot be used for scanning the horizon by rotating the upper mirror because of the image rotation which accompanies such movement. In the panoramic sight (Fig. 5), this difficulty is overcome by providing the system with a dove prism which rotates at half the angular speed of the right-angle prism through the action of a differential gear linkage. The combined inversions of the dove prism and the amici prism at the bottom completely compensate for those of the telescope system, while the relative motions of the right-angle prism and the dove prism maintain the image erect during scanning. See PRISM, OPTICAL.

**Submarine periscope.** In this device, it is necessary to employ a telescope system having a wide field of view and uniform illumination across the field which can be fitted into a long, narrow tube whose length-to-diameter ratio may be 50 or greater. This is achieved by utilizing a plurality of lenses so spaced along the length of the tube as to cause the incoming principal rays from the edge of the field to be deviated from side to side within the tube. In general, the greater the number of lenses, the wider the field of view. One example of the periscopic relay train is shown in Fig. 6 and employs six lenses with three inversions. The typical submarine periscope (Fig. 7) may be considered to be a pair of telescopes facing each other, with such a relay train between. The usual magnification of

submarine periscopes is 6, although some U.S. Navy periscopes have dual magnifications of 6 and 1.5, the latter being achieved by inserting an inverted Galilean telescope into the optical path before the top objective.

The submarine periscope can be provided with a built-in rangefinder for fire-control purposes. A conventional coincidence or split-field type of rangefinder may be attached either vertically or horizontally to the upper end of the periscope, the objective of which receives an image from each of the entrance windows of the rangefinder. See RANGEFINDER, OPTICAL; SUBMARINE.

**Other types.** Various modifications of the basic optical systems described here are employed as viewing periscopes in military aircraft and as viewing devices in particle accelerators and nuclear reactors. The cystoscope and endoscope are slender, sometimes mechanically flexible, periscopes used for visual examination and photography of body cavities inaccessible to direct observation.

[J.K.K.]

**Bibliography:** C. H. v. Hofe, *Fernoptik*, 1941; D. H. Jacobs, *Fundamentals of Optical Engineering*, 1943; A. König, *Die Fernrohre und Entfernungsmesser*, 1937; L. C. Martin, *Technical Optics*, vol. 2, 1950.

## Perissodactyla

An order of hoofed mammals in which the axis of the foot passes through the middle toe, often called the odd-toed ungulates. The order includes the horses, tapirs, rhinoceroses, and their extinct relatives. Perissodactyls first appeared in the early Eocene, reached the height of their evolutionary history in the Oligocene when they were the dominant ungulates in most of the world, and have been on the decline ever since. The order is represented by two principal evolutionary lines. The Hippomorpha include the horses and three other groups now extinct: the paleotheres, titanotheres, and chalicotheres. The Ceratomorpha include the tapirs, rhinoceroses, and several families that became extinct in the late Eocene and Oligocene: the isectolophids, lophiodontids, helaletids, hyrachyids, hyracodontids, and amynodonts. See EUTHERIA; PERISSODACTYLA FOSSILS.

[D.D.D.]

## Perissodactyla fossils

The odd-toed ungulates, hoofed herbivores represented today by the horse, tapir, and rhinoceros families, have a fossil record that includes nine additional families. This record forms one of the most completely known chapters in the history of the Mammalia. Horses in particular provide a record that in richness of detail has produced the most compelling paleontological evidence for evolution as a natural process. The origin of the perissodactyls, however, is still obscure, although it is clear that the Condylarthra provide ideal structural antecedents. Perissodactyls first appear in the early Eocene of North America and Europe and evolve



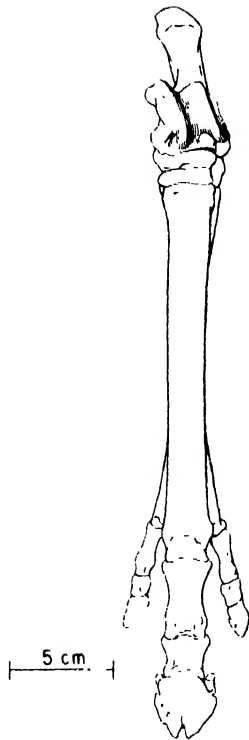


Fig. 1. The odd-toed, or mesaxonic, foot of the perissodactyls. The left hind foot of *Merychippus sejunctus*, a late Miocene horse, showing the astragalus (shaded) with its single pulleylike joint. (After H. Osborn, 1918)

rapidly, reaching a peak of diversity by late Eocene time. Their later evolution, possibly conditioned by rising artiodactyl competition, involves decreasing diversity and emphasis on certain of the better-adapted lines. See ARTIODACTYLA FOSSILS: CONDYLARTHRA: EVOLUTION, ORGANIC.

The order is characterized by emphasis on the third digit as the principal weight bearer; the first is always and the fifth is usually suppressed with varying collateral reduction of the second and fourth digits yielding tetra-, tri-, and monodactyl conditions. The astragalus has a single pulleylike joint articulating with the tibia (Fig. 1). The tarsal articulation is flat; the femur has a third trochanter. The cusps of the cheek teeth tend to be joined by ridges (lophodonty) and the premolars become progressively molariform.

Two suborders have been recognized: the Hippomorpha, including the horses, rhinoceroslike brontotheres and the peculiar, clawed chalicotheres; and the Ceratomorpha, including the tapirs and rhinoceroses. Both suborders diverged from a common, but as yet unknown, ancestry in the later Paleocene. After the early Eocene their evolution proceeded along separate lines. The hippomorphs favored the development of a W-shaped ectoloph and strong styles on the upper cheek teeth (Fig. 2a), while the ceratomorphs retained a simple linear ectoloph and inconspicuous styles (Fig. 2b). Horses and rhinoceroses evolved high-crowned teeth in the late Tertiary, apparently in

response to the spread of grassland environments at that time.

**Hippomorpha.** Horses (North America and Europe) and their close allies, the palaeotheres (Europe), appeared in the early Eocene, while their contemporaries the brontotheres seem to have been restricted to North America. The rapid evolution of the brontotheres was remarkable, for by late Eocene time, this group had reached its greatest diversity, producing some of the largest land mammals then known. No more horses reached Europe until the Miocene, but brontotheres apparently reached both Europe and Asia by late Eocene time and lingered there until the middle Oligocene. Extinction of the brontotheres had already taken place in North America by the close of the early Oligocene. The chalicotheres first appeared in North America and Eurasia in the late Eocene but did not survive the middle Miocene in North America. They apparently reached Africa in the latest Cenozoic and survived there sometime after their extinction in the Pliocene of Eurasia. The main stream of horse evolution clearly took place in North America from which Miocene, Pliocene, and Pleistocene invasions of Eurasia were launched. South America was invaded more than once by horses and tapirs in the late Cenozoic.

**Ceratomorpha.** Primitive tapiroids were common in the Eocene of Eurasia and North America, but the modern family did not enter the record until Oligocene time. Although they inhabited the Northern Hemisphere until the close of the Pleistocene, tapirs survive today only in the Old and New World tropics. The rhinoceroses first appeared in the early Eocene of North America and by Oligocene time inhabited the whole of the Northern Hemisphere. At that time they seem to have attained their maximum development, producing such giants as *Paraceratherium*, the largest terrestrial mammal known. By Miocene time the Eocene lines

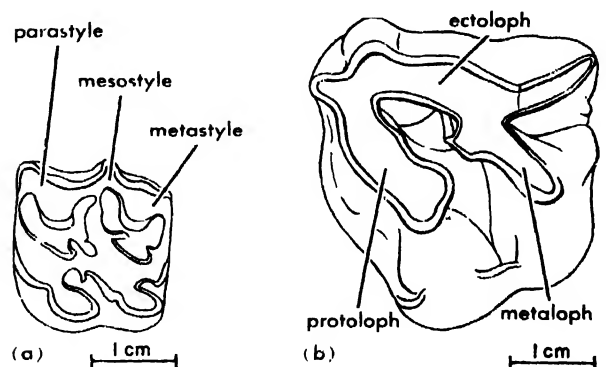


Fig. 2. The first upper molar of representative perissodactyls. (a) A late Miocene hippomorph, the horse *Merychippus sejunctus* (from H. Osborn, 1918). (b) An early Miocene ceratomorph, the rhinoceros *Diceratherium annectens*.

had become extinct and only the modern family remained. At that time the rhinos had reached Africa, and by the close of the Pleistocene they were restricted to that continent and southeast Asia, having vacated their North American birthplace as early as the middle Pliocene. [R.H.T.]

## Peritoneum

The membranous lining of the abdominal cavity, composed mainly of flattened epithelial cells that produce a small amount of watery, or serous, fluid. In the embryo the interior body wall is covered with this membrane which continues over the developing tubular viscera, so that they are suspended and supported by the reflected peritoneum, principally from the posterior body wall. See EPITHELIUM.

As the organs develop, enlarge, and assume their adult form and arrangement, the supporting peritoneum becomes modified, some being lost and other portions becoming thickened, twisted, or otherwise adapting to normal growth. After development is completed those portions which line the interior body wall are called the parietal peritoneum, and the supporting sheets are known collectively as mesentery, although many areas have received specific names. The remaining peritoneum which covers most of the organs is called the visceral peritoneum and this forms the outer layer, or serosa, of the walls of portions of the gastrointestinal tract.

The remaining space, containing a small amount of fluid, between the serosa and the parietal peritoneum is the remnant of the coelom, or body cavity.

In lower vertebrates there is less complexity of development of the viscera, so that the peritoneum in fishes, for example, remains a fairly straight suspensory structure which supports the tubular digestive tract and continues around the body cavity to line the inner walls. [E.G.ST.]

## Peritonitis

Inflammation of the peritoneum, the serous membrane which lines the abdominal cavity and surrounds most of the abdominal organs. The condition may be caused by infectious organisms or foreign substances introduced into the abdominal cavity. The small amount of serous fluid normally present as a lubricant acts as an excellent culture medium for bacterial growth and also as a means of spreading invading materials. The source of such substances or organisms is commonly a gastrointestinal inflammation, especially if perforation has occurred. Appendicitis, peptic ulcer, cancer of the bowel, gallbladder disease, and dysentery are common sources of infection that may produce peritonitis, as well as blood-borne forms of tuberculosis and pneumonia.

Infection may also stem from spread of bacterial organisms from the female organs, the kidneys, the pancreas. Each form of peritonitis may show both common and specific features.

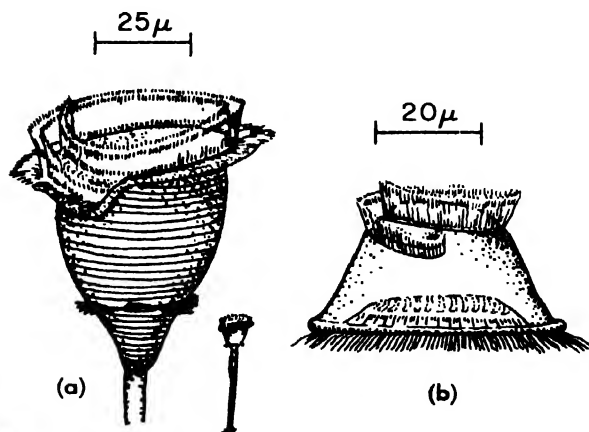
Contamination of the abdominal cavity may occur from penetration of the abdominal wall during an accident or following a wound. Bile, blood, or fluid from a ruptured abdominal cyst or ectopic pregnancy may also induce a peritoneal inflammation.

Where peritonitis occurs, the normally glistening peritoneum becomes dull, the blood vessels engorge, and a fibrin-containing exudate is produced which may later lead to adhesions. A tendency for localization is apparent, with loops of intestine or other organs forming pockets of inflammation.

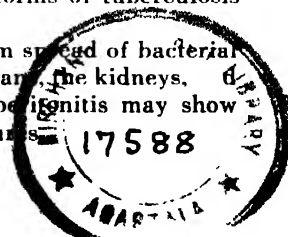
The clinical course is quite variable, depending on the agent involved, the type and severity of reaction, and concomitant disease. See AMEBIASIS; BACILLARY DYSENTERY; GALLBLADDER; PEPTIC ULCER; PNEUMONIA; PREGNANCY, DISORDERS OF; TUBERCULOSIS. [E.G.ST.]

## Peritrichida

An order of the Holotricha composed of a group of unusual-looking ciliates that has excited the curiosity of microscopists for nearly 300 years. Many are sessile and stalked while some form colonies which may reach a large size. A number are attached, as ectocommensals, to a variety of animals and plants. A free-swimming stage in the life cycle, indispensable for distribution, is known as the telotroch. It is a small, mouthless form equipped with a single girdle of posteriorly located locomotor cilia. This is quite unlike the morphology of the mature, sedentary form which is an inverted bell form atop a long stalk. The body is naked of cilia except for conspicuous wreaths of buccal ciliary organelles at the oral end. Much is made, by many protozoologists, of the fact that the adoral zone of membranelles, in this instance, winds counterclockwise toward the mouth. Actually, this is of little real importance, although convenient in taxonomic keys. *Vorticella* (a of illustration) and *Epistylis* are probably the best-known stalked forms. The former is a solitary ciliate, the latter a colony-builder. *Trichodina* (b of illustration) belongs



Peritrichida. (a) *Vorticella*, a stalked peritrich. (b) *Trichodina*, a sessile peritrich.



Re 2212.50  
19:5.67

to the group of mobile peritrichs. Its species are associated with a wide variety of invertebrate and vertebrate hosts on which its actions range from those of a harmless commensal to a pathogenic parasite. See HOLOTRICHA. [J.O.C.]

## Periwinkle

A name applied to various land, fresh-water, and marine snails, but properly denoting the small marine snail, *Littorina littorea*, the typical species of the family Littorinidae, class Gastropoda, phylum Mollusca.

The periwinkle is a native of Europe, where it is a staple food animal. Thousands of tons are sold annually in England alone. It has been introduced at various points along the Atlantic Coast of North America, and is now abundant from Labrador southward to Cape May, New Jersey. There is a limited use of it for food in the United States.

This animal is also of some importance as a fish food, and is sometimes used for fish bait. It has become especially abundant along the Maine coast, where large numbers are exposed by low tide on rocks, wharves, seaweeds, marsh grass, or in ditches and tide pools.

This snail may grow to 1 in. in length and has a variously colored, roughened shell, broad and rounded with 6 or 7 whorls and an acute apex. The outside of the shell is glossy and is marked with dark bands on a background of yellow, brown, olive, red, or black. The inside of the shell varies from white to brown. The periwinkle's foot is divided longitudinally, so that it swings from side to side as it moves. Its head projects from the shell and is equipped with two conical tentacles, with eyes at their bases. The shell is equipped with a horny operculum.

Other species of the genus *Littorina*, all commonly called periwinkles, occur on all coasts of the United States.

In fresh water any small, abundant snail is likely to be called periwinkle. The name is also an accepted common name for a plant, *Vinca minor*, which is grown widely as a ground cover. See GASTROPODA; SNAIL. [J.D.B.]

## Perlite

A natural glass with abundant spherical or convolute cracks which cause it to break into small pearl-like masses or "pebbles," usually less than a centimeter across. It is commonly gray or green with a pearly luster due to reflections from the thin air films formed along the perlitic fractures. Perlitic cracks are not necessarily confined to perlite but appear sporadically in most natural glasses. Glass is formed by rapid cooling of molten rock material (lava), and the cracks are generally believed to develop by contraction during cooling. The water content of perlite is commonly 3-4% by weight. Most of this moisture is believed to have been absorbed by the glass from its surroundings. Some studies suggest that perlitic cracks may form in response to this hydration.

Under heat treatment (about 1500-2000°F) the contained moisture forms tiny steam bubbles in the softened glass, and the perlite is "popped" or exploded to roughly 15 or 20 times its original volume. Thus, the material is excellent for light-weight aggregate, insulation, fillers, and filters. Notable deposits are worked in California and New Mexico. Perlite may constitute major portions of lava flows or occur in small intrusions (dikes). See IGNEOUS ROCKS; LAVA; VOLCANIC GLASS. [C.A.CA.]

## Permafrost

Perennially frozen ground occurring wherever the temperature remains below 0°C for several years whether the ground is actually consolidated by ice or not and regardless of the nature of the rock and soil particles of which the earth is composed. Perhaps 25% of the total land area of the earth contains permafrost; it is continuous in the polar regions and becomes discontinuous and sporadic toward the equator. During glacial times permafrost extended hundreds of miles south of its present limits in the Northern Hemisphere.

Permafrost is thickest in that part of the continuous zone that has not been glaciated. The maximum reported thickness, about 620 m, is at Nordvik in northern Siberia. Average maximum thicknesses are 275-400 m in northern Alaska and 300-500 m in northern Siberia. In Alaska the general range of thickness in the discontinuous zone is 50-150 m and in the sporadic zone less than 30 m. See GLACIAL EPOCH.

Temperature of permafrost at the depth of no annual change, about 10-30 m, generally is below -5°C in the continuous zone, between -1 and -5°C in the discontinuous zone, and above -1°C in the sporadic zone. Temperature gradients vary horizontally and vertically from place to place and from time to time.

Ice is one of the most important components of permafrost, being especially important where it exceeds pore space. Physical properties of permafrost vary widely from those of ice to those of normal rock types. The cold reserve, that is, the number of calories required to bring the material to the melting point and melt the contained ice, is determined largely by moisture content. Ice occurs as individual crystals ranging in size from less than 0.1 mm to at least 70 cm in diameter. Aggregates of ice crystals are common in dikes, layers, irregular masses, and ice wedges. These forms are derived in many ways. Ice wedges characterize fine-grained sediments in continuous permafrost and join to outline polygons. Microscopic study of thin sections of ice wedges reveals complex structures of which some reflect accumulation of ice in seasonal contraction cracks.

Permafrost develops today where the net heat balance of the surface of the earth is negative for several years. Much permafrost was formed thousands of years ago but remains in equilibrium with present climates. Permafrost eliminates most ground-water movement, preserves organic remains,

restricts or inhibits plant growth, and aids frost action; it is also one of the most important factors in engineering problems in the polar regions. See FROST ACTION; MASS WASTING; SOLIFLUCTION.

[R.F.B.]

**Bibliography:** R. F. Black, *Permafrost*—A review, *Bull. Geol. Soc. Am.*, 65:839–856, 1954; S. W. Muller, *Permafrost*, USGS and U.S. Corps of Engineers, Strategic engineering study, Spec. Rept. 62, 1945 (reprint 1947).

## Permalloy

An alloy of iron and nickel, with or without small or moderate additions of other elements. Permalloy is of scientific and technical interest because it has unusually high magnetic permeability.

The nickel content varies from about 40 to 80%. The most commonly added element is molybdenum (4–5%), but similar amounts of chromium and copper are also used; these increase both the permeability and the resistivity. The highest permeability is obtained in Supermalloy, in which are combined the beneficial effects of alloy addition (5% Mo), high purity (aided by heat-treatment in pure hydrogen at 1300°C), and optimum cooling rate (1–5°C/min) through the critical temperature of ordering (about 500°C).

The binary alloys containing from 50 to 80% nickel are susceptible to magnetic anneal; that is, their magnetic properties are drastically changed if they are cooled from about 600 to 400°C in the presence of a magnetic field of a few oersteds. The maximum permeability is thereby raised and the hysteresis loop becomes rectangular in form.

For the compositions and properties of the most used permalloys, see MAGNETIC MATERIALS; see also PERMEABILITY, MAGNETIC.

[R.M.BO.]

## Permanganate

The deep purple anion,  $\text{MnO}_4^-$ , which is derived from permanganic acid,  $\text{HMnO}_4$ . Although the parent acid is stable only in dilute aqueous solution, the salts are well characterized. The permanganates resemble the perchlorates in their oxidizing properties and solubility of both the heavy metal and alkaline-earth metal salts.

Potassium permanganate, the most common permanganate, is produced from a mixture of potassium hydroxide and manganese dioxide which has been oxidized by potassium chlorate, chlorine, or ozone. Permanganates are used as disinfectants, oxidizing agents, wood preservatives, and bleaching agents. See MANGANESE COMPOUND; OXIDIZING AGENT.

[E.E.WR.]

## Permeability, magnetic

A factor, characteristic of a material, that is proportional to the magnetic flux density (magnetic induction)  $B$  produced in the material by a magnetic field divided by the intensity of the field  $H$ . Permeability is usually represented by the Greek letter  $\mu$ .

**Absolute permeability.** Consider a solenoid that has been bent into a circular form so that the ends

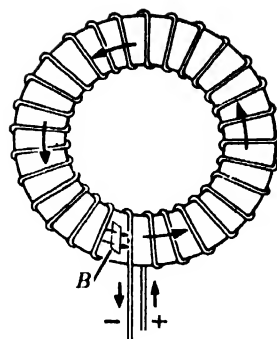


Fig. 1. A solenoid wound in the form of a toroid.

are joined together. This winding is called a toroid (Fig. 1). For a closely wound toroid, almost all the flux is in the interior. The intensity of the magnetic field inside the toroid is

$$H = \frac{NI}{l}$$

where  $l$  is the mean circumference of the toroid and  $I$  is the current in the toroid coil. See MAGNETIC FIELD.

The flux density  $B$  within the toroid is found from Ampère's law to be

$$B_0 = \mu_0 \frac{NI}{l}$$

if the toroid is in empty space (see AMPÈRE'S LAW). Then

$$\mu_0 = \frac{B_0}{H}$$

If a medium takes the place of the empty space within the toroid, the value of  $B$  changes for the same value of  $H$ . The ratio of  $B$  to  $H$  is called the absolute permeability of the medium.

$$\mu = \frac{B}{H}$$

Since the mks unit of  $B$  is the weber per square meter, and the corresponding unit of  $H$  is the ampere per meter, the mks unit of permeability is the weber per ampere-meter. A second unit is the henry per meter. That these two units are equivalent is seen from the relationship  $L = N\Phi/I$ . Since the inductance  $L$  is in henrys when the flux  $\Phi$  is in webers and  $I$  is in amperes, the henry is a weber per ampere. Thus, a henry per meter is the same as a weber per ampere-meter. See INDUCTANCE.

**Relative permeability.** It is convenient to define another quantity, called the relative permeability,  $\mu_r$ , as the ratio of the permeability  $\mu$  of the material to the permeability  $\mu_0$  of empty space.

$$\mu_r = \frac{\mu}{\mu_0}$$

Relative permeability is a pure number, and independent of the system of units used. The permeability of free space has the numerical value of  $4\pi \times 10^{-7}$  henry per meter. See ELECTRICAL UNITS.



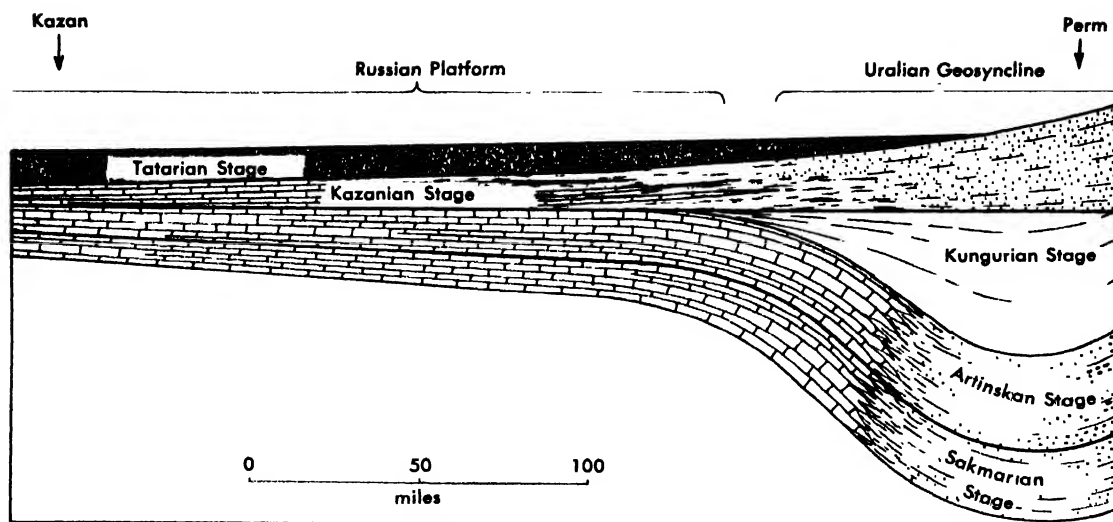


Fig. 1. East-west section of the Permian System in its type region.

about 1920, and it has now become the standard section for North America and, to a considerable degree, for the world. It is now divided into four series as follows: Ochoan series, about 4500 ft; Guadalupian series, about 3000 ft; Leonardian series, about 3000 ft; and Wolfcampian series, 1500+ ft.

The Wolfcampian correlates closely with the Sakmarian stage of the type section in Russia and, like the latter, is the Zone of *Pseudoschwagerina*; the Leonardian correlates with the Artinskian but has a larger and more varied fauna; the Guadalupian cannot be correlated in detail with the upper part of the Russian section, because the latter is largely unfossiliferous. The Ochoan is entirely unfossiliferous except for a thin zone near the top (in the Rustler dolostone) which contains productid (brachiopods) and a few other types of Paleozoic invertebrates.

**Leonardian time.** Facies changes in this region are spectacular. During Early Permian time (Wolfcampian Epoch), the region was a broad, shallow marine basin in which a rich and varied fauna thrived. By Leonardian time three distinct basins (Fig. 3) were subsiding more rapidly than the surrounding area, which became a broad shelf occupied by wide shallow lagoons. Light-colored, fossiliferous limestones (Victorio Peak limestone) then accumulated on the shelves, while black limestone and black shale accumulated in the basins. Evidently the threshold to the basins, which was somewhere in Mexico, was then so shallow that wa-

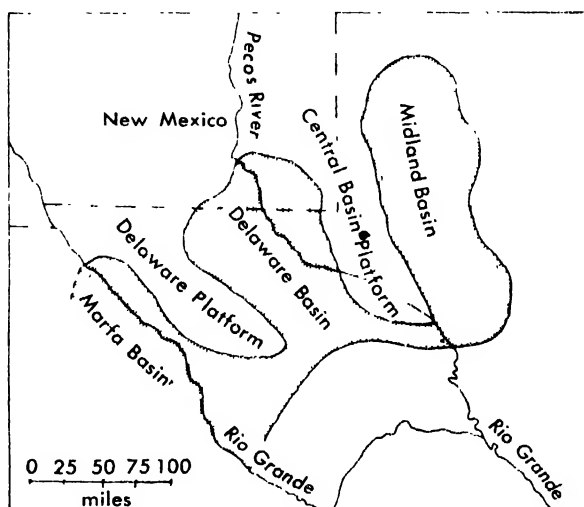


Fig. 3. Permian Basins of West Texas. (After P. B. King)

ter in the basins was density-stratified and the bottom was stagnant and foul, so that almost no benthonic organisms could survive. The Bone Spring black limestone is generally barren of fossils.

**Guadalupian time.** During Guadalupian time the basins continued to deepen, and as the climate became strongly arid, surface water flowed radially out of the basins onto the platform to replace the water lost by evaporation in the shallow lagoons. As a result, a narrow limy bank grew up along the margins of the basin to form the great Capitan Reef. With this development, three strongly contrasted facies accumulated simultaneously: (1) basin or pontic deposits under normal marine conditions, (2) reef and reef talus, and (3) back-reef or lagoonal deposits. Figure 4 shows the complex relations within the Leonardian and Guadalupian deposits along the face of the Guadalupe Mountains. Massive deposits of reef talus were derived from the growing front of the reef. These dip steeply into the basin, become finer down dip, and grade

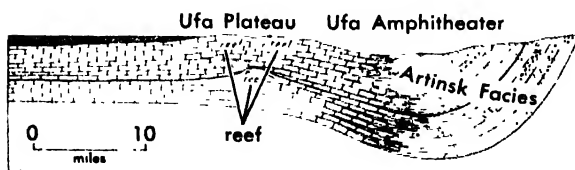


Fig. 2. Idealized section showing relations of the Artinsk detritals to the limestones of the Ufa Plateau.

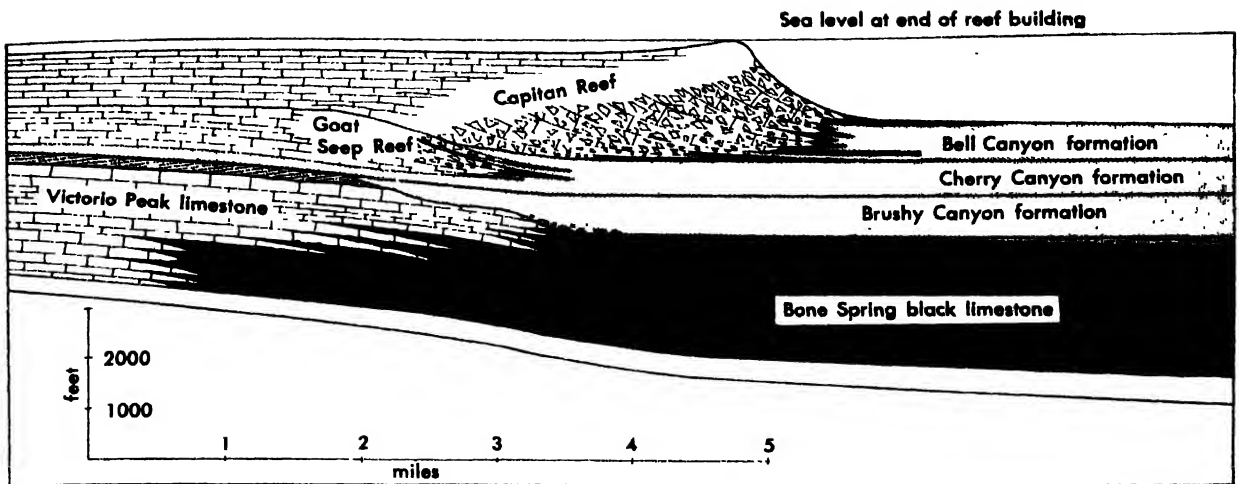


Fig. 4. Section across the Capitan Reef Complex. (After P. B. King)

out into thin tongues of calcarenite. The back-reef deposits are calcareous for distances varying from 1 to 5 miles. The deposits then grade rapidly into gypsiferous shales and anhydrite. Gray sands intertongue from the landward margins of the lagoon. Farther back the sands pass into redbeds.

*Ochoan time.* Finally, during Ochoan time the marine water withdrew entirely into the basins and, under intensely arid conditions, became a dead sea in which enormous deposits of anhydrite and, later, of halite were precipitated. Interbedded with the salt in the center of the Delaware Basin are several lenses of potash salts—sylvite, carnallite, and polyhalite.

Probably no other great reef complex is so well exposed or has been so intensively studied as that of Capitan Reef. Cenozoic faulting uplifted the Guadalupe Mountains, but immediately to the west it produced the Salt Flat graben. Thus, a great natural section across the reef complex was exposed. The Capitan Reef was essentially a slender limy bank along the margin of the platform, and much of the carbonate may have been precipitated biochemically. Much of it is sparingly fossiliferous, but in places it is filled with fossils which locally accumulated in a typical reef environment.

From study of deep well cores, the West Texas formations are known to extend eastward under the Staked Plains to central Texas where they reappear at the surface. Thence their outcrops stretch northeastward across Oklahoma and Kansas. The Wolfcamp equivalents are largely of marine origin in central Texas and Kansas. Marine conditions persisted well into Leonardian time in central Texas, but in Kansas a very large salt deposit (in the Wellington shale) is followed by a thick redbed sequence. By Guadalupian time most of the deposits in central Texas were nonmarine redbeds with several thin marine dolostones intertonguing from the west. In Oklahoma all but the lower part of the Wolfcampian is in the redbed facies because of the local influence of the Oklahoma mountains.

**Other regions.** Permian deposits are thick and widespread in many other parts of the world, particularly in South Africa, Australia, India, China, Indochina, and Japan, and probably also in South America.

*South Africa.* The Permian of South Africa is entirely nonmarine and is notable for its extensive glacial deposits as well as its fossil reptiles. Here the section begins with the widespread Dwyka tillite, followed by the Eccra series and the lower part of the Beaufort series. These three constitute the lower part of the Karoo System. The Dwyka tillite is discussed below under Permian glaciation.

Both the Eccra and Beaufort series are very thick in the old geosyncline which paralleled the southern coast of Cape Colony, but each thins northward and spreads widely over the southern part of Africa. The Eccra series consists of dark shales and sandstones and includes the Coal Measures of Africa. The Beaufort series was spread over the Eccra series after an interval of erosion. It differs from the Eccra series in having interbedded yellowish sandstones, some members of red shale, and distinctive reptilian faunas.

*Australia.* The Australian continent was largely emergent during Permian time except for a series of marginal basins along both the western and eastern coasts. These basins subsided independently; although the Permian deposits are thick, it is impossible to treat them adequately within the limits of this article. The most significant fact is that the Permian here, as in South Africa, begins with widespread glacial deposits. In the Gascoyne Basin of northwest Australia these include tillite which represents a ground moraine resting on an old erosion surface. In most of western Australia, however, the glacial deposits appear as boulder beds intercalated in marine sandstone or shale, a clear indication that the boulders were dropped from floating ice. In the western basins, interbedded marine zones in the Lyons sandstone bear the ammonite *Metalegoceras jacksoni* which dates the deposits as Sakmarian or early Artinskian. After deposition



of thick glacio-marine deposits of this sort, each of the western basins was occupied by a warmer marine incursion in which widespread and fossiliferous, but rather thin, calcareous formations accumulated (Fossil Cliff limestone of Irwin River Basin, Callytharra limestone of Gascoyne and Minilya basins, and Nura Nura limestone of the Kimberley District). Their faunas are of Artinskian age. In both the Irwin River and Gascoyne basins, glacial erratics recur for some distance above this limestone horizon, which proves that glaciation persisted here at least into early Artinskian time.

In southeastern Australia where Permian rocks are very thick they have been subdivided into four major units, a lower and an upper marine unit and Lower and Upper Coal Measures. Glacial erratics occur in both of the marine units.

The Permian strata in Germany comprise two major units. The lower is the Rothliegendes sandstone, consisting of nonmarine redbeds; the upper is the Zechstein formation consisting of marine limestones and shales which grade upward into salt and gypsum. The potash deposits at Stassfurt lie within the upper part of the Zechstein. The magnesian limestone of England is a thin western tongue of the Zechstein limestone.

**Permian glaciation.** Early in Permian time, large areas in the southern continents were covered by glacial ice. The widespread Dwyka tillite indicates that most of Africa south of latitude 23°S was ice-covered. It includes abundant striated boulders and in places rests on a spectacularly grooved and striated floor. In its northern outcrops (northern Karoo, Natal, and Zululand), it is a typical ground moraine resting on an undulating pre-Permian surface; but farther south it thickens greatly and grades into a glacio-marine deposit formed of material dropped from floating icebergs. Orientation of glacial striae and distribution of boulders from known source areas indicate that the ice moved westward into the province of South Africa and generally south from the Transvaal. N. Boutakoff (1940) has reported that tillite (presumably Dwyka) is widespread in the Congo Basin, even within 4° of the Equator.

Glacial deposits are also widespread in western and southern Australia and in Tasmania. In western Australia such deposits originally covered an area of about 200,000 square miles. Locally, in the Canning Basin the glacial deposits are typical ground moraine resting on a striated floor, but for the most part they are glacio-fluvial in western Australia and occur within a thick marine sequence (the Lyons formation of the Irwin and Canning Basins, and the Kungangie and Grant Range formations of the Kimberley District).

Permian glacial deposits are also widespread in South America (in Uruguay, in the Precordillera of northern Argentina and Bolivia, and in southern Brazil). An extensive ice sheet in India is recorded by the Talchir tillite of the Salt Range and central India (Rewah Province). The ice sheet is believed to have stretched for some 600 miles from east to

west and 1000 miles from south to north and to have moved northward into the Salt Range region.

Curiously, no glaciation is known to have occurred elsewhere in the Northern Hemisphere.

**Date of ice age.** In each of the regions mentioned above, the glacial deposits are at the base of the Permian section and the immediately overlying fossiliferous formations are of Artinskian age; hence it appears that the glaciation occurred early in the period, during either Sakmarian or early Artinskian time. In India the Talchir tillite is overlain by the Lower Productus limestone which bears a prolific marine fauna including fusulines of Artinskian age. In South Africa the tillite is succeeded by the Ecce formation which bears vertebrate fossils of mid-Permian age, and in western Australia the glacio-marine deposits include the ammonite, *Metalegoceras jacksoni*, which is correlated with the Lower Permian or early Artinskian faunas of Timor.

The reptiles in the Ecce formation of South Africa indicate a mild climate, as do the marine faunas of the Lower Productus limestone of India. It appears therefore that the glacial episode occupied but a relatively short part of Permian time. The glacial deposits of South America are not well dated by fossils and have been classified by most South American geologists as Upper Carboniferous; but because their occurrence so closely resembles that of Australian glacial deposits, it seems highly probable that they are of Early Permian age.

**The *Glossopteris* flora.** Throughout the glaciated regions of the Southern Hemisphere the nonmarine Permian formations are commonly characterized by the tongue ferns, *Glossopteris* and *Gangamopteris*. In South Africa *Glossopteris* has been found between the glaciated floor and the Dwyka tillite, and in numerous places in India, South Africa, and Australia its distinctive spores have been found in the tillite. It is therefore believed to have been adapted to a cold climate. Unfortunately, the biologic relations of these peculiar plants are still uncertain.

**Permian deserts.** Desert conditions probably were more widespread in the Permian than at any other time before the late Cenozoic. The vast deposits of salt and anhydrite in the Permian Basin of West Texas and New Mexico, the salt of Kansas, and the extensive deposits of dune sand in the eastern part of the Colorado Plateau indicate a great arid basin in the west-central part of the United States. Ralph King has estimated that, if it required 300,000 years to precipitate the salines of the Ochocoan series, the rate of evaporation over the entire basin must have averaged 9.5 ft per year, which is only about 2 ft less than it is in the modern Death Valley of California.

Similar conditions must have obtained during deposition of the Kungurian salts in the Permian Basin of the Soviet Union. Western Europe was also the scene of great aridity, as indicated by the thick salt deposits at Stassfurt, Germany, and by widespread dune sands in Germany and England.



The three greatest known deposits of potash salts lie within the Permian System, one in west Texas, one about Solikamsk in the Permian Basin of the Soviet Union and one about Stassfurt, Germany. Although these basins were of regional extent, they were local in the world scene, and in some regions (notably South Africa, Australia, and south China) the Permian System includes much coal and abundant evidence of humid conditions.

**Permian orogeny.** The Permian was a time of continental uplift and of widespread orogeny. At this time the Appalachian Mountain system was formed and the Uralian geosyncline of the Soviet Union was folded and uplifted into a great mountain chain. In Europe the Variscan Alps stretched from southern England across central Germany and from Normandy into the central Massif of France and thence northeastward through the area of the modern Schwarzwald and northeastward through the Erzgebirge to beyond Vienna. The folding of the Variscan Mountains occurred in three stages, the first in the Lower Carboniferous, the second in the Upper Carboniferous, and the final movement in Permian time. There was Permian orogeny also in the area of the modern Kuen-Lun Range in the northern flank of the Himalayas. In none of these regions is the uplift within the Permian closely dated, but before the end of the period the continents were almost completely emergent; the youngest known Permian deposits are generally nonmarine. This may, in part at least, account for the profound change in so many groups of animals and plants at the close of the Paleozoic Era. See OROGENY.

**Permian life.** The invertebrate faunas of the Permian developed from those of the Pennsylvanian with gradual change and marked specialization. Brachiopods, bryozoans, and fusulines and goniatites continued to dominate the faunas, but at the close of the period each of these groups suffered a great decline. Among the brachiopods the Productidae were especially varied and gave rise to highly specialized offshoots such as the leptodids, the richtofenids, and the scacchinellids, all of which were associated with reef facies. By the end of the period all the productids were extinct. The bryozoans were prolific and highly varied, but by the close of the period two of the chief Paleozoic orders, the Trepostomata and the Cryptostomata, were extinct. The fusulines and neoschwagerines were extraordinarily abundant until near the end of Permian time and reached their maximum size, but none survived beyond this period. Goniatites expanded rapidly into several families, in some of which complex sutures reached the typical ammonitic stage; but late in the period they suffered near extinction; only a few genera of two families survived to start a spectacular new evolution in the succeeding Triassic Period. See TRIASSIC.

On land the insects showed a great advance over those of the Coal Measures, and several of the modern orders emerged, among them the Mecoptera,

Odonata, Hemiptera, Copegnathia, Hymenoptera, and Coleoptera. Extensive insect faunas have been found in the Lower Permian rocks of Kansas and Oklahoma, the Permian Basin of the Soviet Union, and the Upper Permian of Australia. See INSECTA FOSSILS.

Land plants displayed at first a gradual, and eventually a profound, change as the dominant lowland plants—the lepidodendrons, sigillarias, and cordaites—of the moist coal swamps declined, and the conifers advanced to a dominant position.

Of the vertebrates, the labyrinthodont amphibians were common and varied, but the reptiles showed the most significant advances. Reptiles have been found in abundance in the lower half of the system in Texas, throughout most of the system in the Permian Basin of the Soviet Union, and in the Karoo series of South Africa. Nearly all of the Permian reptiles were short-legged sprawlers. Of several orders, the most significant were the Theriodonts or mammal-like reptiles that foreshadowed the advent of the mammals. These reptiles carried their bodies off the ground and walked or ran like a mammal instead of sprawling. Their teeth became specialized incisors, canines, and jaw teeth as in the mammals—and all the elements of the lower jaw except the mandibles showed progressive reduction. Most of the known theriodonts are from South Africa and the Soviet Union, but a few typical genera have been found in Permian beds of the Cordilleran region in North America. See THERAPSIDA; see also BLACK SHALE; EVAPORITE (SALINE); FACIES (GEOLOGY); GEOSYNCLINE; REDBED. [C.O.D.]

*Bibliography:* A. L. DuToit, *Geology of South Africa*, 3d ed., 1954; *Geologisches Steuere, USSR*, 1:345–372, 1958; P. B. King, *Geology of the Southern Guadalupe Mountains, Texas*, USGS Prof. Paper 215, 1948; N. D. Newell et al., *The Permian Reef Complex of the Guadalupe Mountains Region, Texas and New Mexico*, 1953; K. Teichert, Upper Paleozoic of western Australia: correlation and paleogeography, *Bull. Am. Assoc. Petrol. Geologists*, 25:371–415, 1941.

## Permittivity

The permittivity  $\epsilon$  of a material medium is related to the permittivity  $\epsilon_0$  of empty space by the equation  $\epsilon = k\epsilon_0$  where  $k$  is the relative dielectric constant of the medium. The permittivity of empty space  $\epsilon_0$  has the value  $8.85 \times 10^{-12}$  coulomb<sup>2</sup>/newton-meter<sup>2</sup>. See CAPACITOR; COULOMB'S LAW; DIELECTRIC CONSTANT; ELECTRICAL UNITS. [R.P.W.]

## Perovskite

A mineral with composition  $\text{CaTiO}_3$ . Morphologically, perovskite crystals are isometric but the internal symmetry is lower, probably monoclinic. The hardness is 5.5 (Mohs scale) and the specific gravity is 4.0 or higher in some varieties because of the presence of cerium or columbium. The luster is metallic in black material but adamantine in red-to-yellow varieties. Perovskite occurs as an acces-

sory mineral in basic igneous rocks, particularly those containing nepheline or leucite. It is also found in basic pegmatites and in limestone at the contacts of basic or alkaline intrusions. Perovskite may form as an alteration product of sphene and ilmenite. *See* ILMENITE; SPHENE; TITANIUM.

Barium titanate, which is structurally similar to perovskite, is important as a component in modern electronic equipment, such as transducers. *See* BARIUM TITANATE. [C.S.HU.]

## Peroxide

A chemical compound which contains the peroxy ( $-O-O-$ ) group, which may be considered to be a derivative of hydrogen peroxide ( $HOOH$ ). An organic (or inorganic) peroxide is one in which some organic (or inorganic) substituent has replaced one or both hydrogens. Peroxides are used in such diverse reactions as oxidation, synthesis, polymerization, and oxygen generation. Inorganic peroxides include persulfates, hydrogen peroxide ( $H_2O_2$ ), sodium peroxide, bivalent metal peroxides, and  $H_2O_2$  addition compounds. Organic peroxides include per(oxy)acetic acid, dibenzoyl peroxide, and cumene peroxide.

**Inorganic peroxides.** Peroxydisulfates, familiarly called persulfates, are produced by electrolytic oxidation of aqueous sulfuric acid or ammonium bisulfate. Persulfuric acid ( $H_2S_2O_8$ ) is not used commercially as such. Ammonium persulfate, a white solid, is used as a polymerization catalyst, dyestuff oxidant, metal etchant, and laboratory oxidant. The potassium salt is used extensively in the manufacture of styrene-butadiene synthetic rubber; smaller quantities are used in hair bleaches.

Peroxymonosulfates, also called monopersulfates, are salts of peroxymonosulfuric acid (Caro's acid). The latter, made from  $H_2O_2$  and sulfuric acid, is a powerful oxidant and bleach; these properties are shared by the salts. The acid can be used for making wool resistant to shrinkage; the salts are effective bleaching agents for domestic laundering.

Peroxydiphosphates, analogous to peroxydisulfates, have been prepared by electrolytic oxidation of concentrated phosphate solutions; no major technical uses have been reported. Peroxydicarbonates have also been made electrolytically.

**Hydrogen peroxide.** The most widely used peroxy compound is hydrogen peroxide, a waterlike liquid manufactured as aqueous solutions of 35–90%  $H_2O_2$  by weight; essentially anhydrous  $H_2O_2$  has recently become commercially available. Annual production in the United States exceeds 50,000,000 lb (100% basis). Selling price in commercial quantities is about 50 cents per lb of contained  $H_2O_2$ .  $H_2O_2$  is not combustible; water is a safe diluent and coolant. With organic compounds,  $H_2O_2$  can form detonable mixtures; industrial processes guard against their generation. Directions for safe handling and storage of  $H_2O_2$  are available from producers, government agencies, and trade associations.

Hydrogen peroxide is manufactured by electrolytic and organic oxidation processes. The former involves electrolytic production of the peroxydisulfate intermediate, followed by steam hydrolysis to  $H_2O_2$ , with regeneration of the original sulfuric acid or ammonium bisulfate raw materials. One organic process uses an anthraquinone dissolved in organic solvents. The quinone is catalytically hydrogenated to the hydroquinone; subsequent aeration of the latter regenerates the quinone, with simultaneous formation of  $H_2O_2$ . The  $H_2O_2$  is water-extracted and concentrated; the quinone is recycled for reconversion to hydroquinone. A second organic process uses liquid isopropyl alcohol, which is oxidized at moderate temperatures and pressures to  $H_2O_2$  and acetone coproducts. After distillation of the acetone and unreacted alcohol, the residual  $H_2O_2$  is concentrated.

Hydrogen peroxide applications include commercial bleaching, dye oxidation, manufacture of organic and peroxy chemicals, and power generation. Bleaching outlets consume more than one-half of the  $H_2O_2$  produced. These outlets include textile mill bleaching of practically all wool and cellulosic fibers, as well as of major quantities of synthetics, and paper and pulp mill bleaching of groundwood and chemical pulps (*see* BLEACHING). Organic applications include manufacture of epoxides and glycols from unsaturated petroleum hydrocarbons, terpenes, and natural fatty oils. The resultant products are valuable plasticizers, stabilizers, diluents, and solvents for vinyl plastics and protective coating formulations. Production of tonnage organic chemicals may become feasible via low-cost  $H_2O_2$ ; synthetic glycerol production, using captive  $H_2O_2$ , has been planned by one  $H_2O_2$  producer.  $H_2O_2$  outlets for manufacture of peroxides include inorganic compounds such as sodium perborate, and organic compounds such as peracetic acid and dibenzoyl peroxide. Certain peroxides, such as that of sodium, are more economically produced by air oxidation. Power generation applications include use in specialized propulsion units for aircraft, missiles, torpedoes, and submarines. The hot oxygen-steam mixture from catalytically decomposed  $H_2O_2$  powers the feed pumps of many large liquid-propellant rockets.

**Metal peroxides.** Sodium peroxide ( $Na_2O_2$ ), a yellowish powder, is the peroxide produced in second largest amount for direct sale. Annual production in the United States is about 12,000,000 lb. Manufacture is via a two-stage reaction of the elements. The sodium monoxide first formed from the reaction of liquid or solid sodium and dried air in rotary steel burners is converted to the peroxide by additional reaction at 250–400°C. Selling price is about 20 cents per lb. From their respective active oxygen contents,  $Na_2O_2$  and  $H_2O_2$  are competitive in price. Aqueous solutions of  $Na_2O_2$  are essentially equivalent to a mixture of caustic soda and  $H_2O_2$ . Contact of the powder with skin or combustibles must be avoided to minimize the danger of burns or fire. In event of fire, salt or sand in-

stead of water must be used. Major uses of  $\text{Na}_2\text{O}_2$ , as with  $\text{H}_2\text{O}_2$ , are in bleaching processes in the textile and in the pulp and paper industries. In certain areas, the two chemicals are competitive; in others, the choice is dictated by the particular application.

Bivalent metal peroxides that are commercially available include those of barium, calcium, magnesium, strontium, and zinc. Those of barium and strontium can be obtained by roasting the metal in air or oxygen; the others are best made by reaction with  $\text{H}_2\text{O}_2$  of a solution of a salt or a slurry of oxide or hydroxide. Large-scale commercial uses have not been developed. The barium and strontium compounds are used for coloring flames in pyrotechnics. Calcium peroxide is used as a dough conditioner in the baking industry; magnesium and zinc peroxides are used cosmetically as deodorants and antiperspirants.

Inorganic peroxy anion compounds are readily synthesized from  $\text{H}_2\text{O}_2$  and solutions of various metal anions (pertitanates, perchromates). Their importance is chiefly in the analytical area; some have importance as catalysts.

Hydroperoxidates are solid addition compounds of  $\text{H}_2\text{O}_2$  with other materials. Sodium perborate "tetrahydrate" ( $\text{NaBO}_2 \cdot \text{H}_2\text{O}_2 \cdot 3\text{H}_2\text{O}$ ) and "monohydrate" ( $\text{NaBO}_2 \cdot \text{H}_2\text{O}_2$ ) are convenient sources of  $\text{H}_2\text{O}_2$  when dissolved in water. Manufacture is by reactions of borax, caustic soda, and  $\text{H}_2\text{O}_2$ . Perborates are used in sizable amounts in household powdered bleaches for the home laundry, and in the textile industry for oxidation of vat and sulfur dyes on cotton and rayon. The hydroperoxidates of sodium carbonate, sodium pyrophosphate, and urea are also available. Fields of use for the latter compound include cosmetics and photography.

**Organic peroxides.** As a group, organic peroxides are more hazardous than the inorganic peroxides. Many of the former are flammable or detonable, thus restricting their availability. Some are exceptionally stable (di-*tert*-butyl peroxide). Manufacture is chiefly through reaction of the organic substrate with  $\text{H}_2\text{O}_2$ ; air oxidation is also used when feasible.

Peracetic acid [ $\text{CH}_3(\text{C}=\text{O})\text{OOH}$ ], prepared from acetic acid and  $\text{H}_2\text{O}_2$ , is the only organic peracid offered commercially (40% by weight in acetic acid). A manufacturing process involving air oxidation of acetaldehyde has been developed. Peracetic acid, as well as performic or perpropionic, may also be generated at the site from  $\text{H}_2\text{O}_2$ , organic acid, and catalyst. Major applications of peracetic acid are in the synthesis of epoxidized and hydroxylated compounds and as a bactericide, fungicide, and sterilizing agent for processing equipment.

Dibenzoyl peroxide, the most important aromatic acyl peroxide, is a white powder, stable at room temperature and explosible with heating. It is manufactured by reaction of benzoyl chloride and alkaline  $\text{H}_2\text{O}_2$ . Major uses include polymer manufacture (0.1–0.2% in the monomer to initiate the poly-

merization) and flour bleaching (when mixed with phosphates and other ingredients meeting standards for flour-treating formulations).

Cumene hydroperoxide [ $\text{C}_6\text{H}_5\text{C}(\text{CH}_3)_2\text{OOH}$ ], a colorless to pale yellow liquid, produced by air oxidation of isopropyl benzene, is no longer used extensively as a polymerization catalyst. However, its ready cleavage in acid solution to phenol and acetone has rendered it a recent tonnage intermediate in the commercial production of phenol. See ELECTROCHEMICAL PROCESS; OXIDATION PROCESS; OXIDATION-REDUCTION; OXIDIZING AGENT; OXYGEN.

[S.S.N.]

**Bibliography:** W. C. Schumb, C. N. Satterfield, and R. L. Wentworth, *Hydrogen Peroxide*, 1955; A. V. Tobolsky and R. B. Mesrobian, *Organic Peroxides*, 1954.

## Perpetual motion

The historic phrase perpetual motion, or *perpetuum mobile*, entails three concepts, the most common of which is perpetual motion of the first kind.

Perpetual motion of the first kind refers to any mechanism that, once set in motion, will continue to do useful work without drawing on any external source of energy or that, in every cycle of its operation, will produce more energy than it absorbs. Although the quest for a machine or engine with an efficiency exceeding 100% has often been cited as one of the classic follies of science, most scientists from the sixteenth century on regarded it as futile, especially insofar as mechanical devices were concerned. However, it was only with the establishment of the general principle of conservation of energy, in the middle of the nineteenth century, that the possibility of obtaining perpetual motion of the first kind could be completely ruled out. Of the many attempts made to achieve perpetual motion of the first kind (for instance, Figs. 1 and 2), a number are of historical interest because they added materially to empirical knowledge or served to clarify puzzling phenomena.

Perpetual motion of the second kind refers to any engine that will convert heat completely into other forms of energy, thus making it possible, for instance, to draw on the enormous store of internal energy in the atmosphere or sea and convert it completely into useful work. The impossibility of

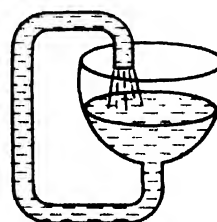


Fig. 1. A perpetual-motion device based on the hydrostatic paradox. The water supposedly would flow continuously because the weight of it in the large vessel exceeds that in the tube.

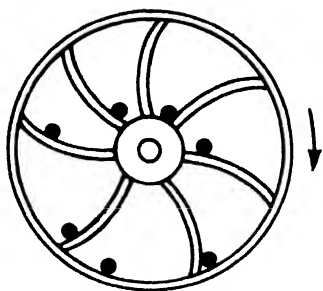


Fig. 2. A seventeenth century proposal for perpetual motion, based on the notion that the torques about the axle, due to the balls rolling in the grooved spokes, will be larger on the descending side of the wheel than on the other side.

having such an engine became evident with the establishment, near the middle of the nineteenth century, of the second law of thermodynamics. It is this principle and not the principle of conservation of energy that rules out perpetual motion of the second kind, for there is nothing in the energy principle that denies the possibility of converting heat completely into work.

The term perpetual motion of the third kind is sometimes used to refer to the continuous motion that a mechanism would have if it were completely freed from the action of all nonconservative forces, that is, forces such as those of friction which extract energy irreversibly from a mechanical system. Experience shows that such forces can be reduced, as by lubrication in the case of friction, but can never be eliminated. If they could, continuous motion of a mechanism would be possible without violation of either the principle of conservation of energy or the second law of thermodynamics. See CONSERVATION OF ENERGY; THERMODYNAMIC PRINCIPLES. [D.E.R.]

**Bibliography:** H. Dircks, *Perpetuum Mobile*, 2 vols., 1861-1870; M. W. Zemansky, *Heat and Thermodynamics*, 4th ed., 1957.

## Perseus

A compact circumpolar constellation of the northern sky, like its neighbor, Cassiopeia, on the east. Both constellations lie in a brilliant part of the Milky Way. The prominent stars in Perseus form the capital script letter *A*. This group is represented by the figure of the hero, Perseus. According to legend he dutifully follows his mother-in-law, Cassiopeia, in her circle about the heavens. The conspicuous curved arc of stars, bright and easy to identify, is commonly known as the Segment of Perseus. Mirfak, a navigational star, lies in the right shoulder. The constellation is noted for its clusters of stars. Just above the head are the famous double clusters in Perseus. Algol, the Demon Star, which is an eclipsing variable, is located in this constellation. See CASSIOPEIA; CONSTELLATION.

[C.S.Y.]

## Persian melon

A long-keeping variety of muskmelon, *Cucumis melo*, of the plant order Campanulales. The fruit is large (6-8 lb), globular, and without sutures; it has dark green skin and thin, abundant netting; the flesh is firm, thick, orange-colored, and sweet, with mild but distinctive flavor. Persian melons require a long warm season, and are highly susceptible to diseases, which are intensified by rain or high humidity. Except in the great central valley of California, the Persian melon is little grown in the United States. The average annual farm value in the United States for the period 1949-1957 was about \$1,000,000. See CAMPANULALES; MELON GROWING; MUSKMELON. [V.R.B.]

## Personality theory

Personality as a technical concept is inferred from consistencies in the behavior of the individual in different situations and over extended periods of time and growth. These consistencies in behavior are usually of three kinds. The first relates to the goals the person strives for and the situations he finds rewarding or threatening. A second form of consistency is the characteristic style with which a person carries out his goal-directed enterprises—his persistency, his manner of learning and thinking, the manner in which he handles obstacles and frustrations, and how he resolves conflicts of need and interest. Finally, there is a consistency in the expression of affect—a person's intensity of feeling, his mood swings, his energy level, the degree to which he is strongly driven toward goals, and the extent to which he is sensitive to internal as compared to external stimulation and demands.

**Behavior consistencies.** To account for these consistencies, psychology postulates certain processes assumed to be operating in the individual in interaction with the environment in which he lives.

**Goal striving.** To account for consistencies in goal striving, for example, various versions of the concept of need, drive, or interest are employed. When the person is found to be striving consistently and persistently for certain end states that appear to bring cycles of behavior to a temporary close once they are attained, one infers that there is a need for the attainment of that end state, whether or not the person can put this need into words. The various needs that are inferred from the behavior of any single individual are not assumed to be independent of one another and are related in certain characteristic ways, for example, when the attainment of one goal is blocked and the person characteristically substitutes another form of goal striving in its place (displacement or substitution). Or needs may conflict with each other so that the behavior related to the attainment of one goal precludes behavior necessary for attainment of another, for example, the conflict between behavior directed toward flight under conditions of fear arousal and the behavior involved in gratify-

ing sexual need, where the arousal of the former renders neurologically impossible the latter activity. The individual's requirements are such that virtually any set of needs can get into a state of conflict should there be overemphasis in the amount of time and activity given to any single one of them, and in consequence the problem of regulation of needs becomes a central one in any theory of personality.

*Style of goal striving.* The second set of postulated characteristics of personality are those having to do with the style or manner in which people carry out their goal-directed activity. In essence, these postulated characteristics are a set of principles having to do with the regulation and synthesis of various goal-striving activities, such as the capacity to delay gratification of one need for the sake of maintaining integration, for example, when hostile impulses are put aside in the interest of a longer-term need to maintain amicable social relations within a group. It is also generally assumed that there is a cost factor in such need regulation—eschewal of one need for gratification of others involves a price in terms of energy expenditure or resultant malaise on the part of the person. In a later section on the concept of defense, the economy of need regulation will be further elaborated.

In addition to these more motivational aspects of regulation and synthesis, other features of personality are usually taken into account that have to do with the person's manner of coping with the problems that arise in goal striving. An example is the handling of frustration—whether it is followed by aggression, by retreat, by realistic problem solving, or by substitution of other more success-assuring activities. Relevant here as well are the skills and techniques that the individual has learned to use in the course of development in the interest of attaining his goals. These are generally treated less as specific skills than as orientations toward problem solving, for example, the tendency toward quick perceptual completion of ambiguous situations, tendencies to be more dependent upon external cues in a situation than upon inner ones. Such orientations are inferred either from factorial studies of specific behavior (discussed later) or by more informal techniques for inferring consistency in the utilization of experience to regulate problem solving.

*Affective reactions.* The third feature of personality deals with the characteristic affective reactions of the person; this area is less highly developed from the point of view of research than the other two. Traditionally, this is taken to be the problem of temperament and it is closely, but not exclusively, linked to emphasis upon constitutional factors. Because inner affective states do not lend themselves to precise measurement, but depend to a large degree on subjective report, temperamental and affective characteristics of the individual are not easily investigated. In consequence, many indirect physiological methods have been developed

for the study of emotional states which do not yield metrics that readily lend themselves to direct translation of the subjective conditions producing them. The classical approach to the problem has been in terms of certain typologies of temperament, such as the one proposed centuries ago by Hippocrates in terms of the sanguine, the melancholic, the choleric, and the phlegmatic, each assumed to be related to the state of the body fluids of human beings. More modern theories, such as those of E. Kretschmer and W. H. Sheldon, are more sophisticated in their use of measurement of bodily indices but use essentially the same mode of inference in arriving at over-all temperamental types. These will be discussed later.

**Study approaches.** The following approaches to personality study are selected as representative of some of the major twentieth-century trends: the psychodynamic approach, the cultural approach, the approach via learning theory, the organismic and self-concept approaches, the factorial trait approach, and the typological-constitutional approach.

*Psychodynamic approach.* The basic tenet of the psychodynamic approach is that the individual's goal striving and motivations are the core from which personality theory must be built. The etiology of motives, their development and transformation, and the resulting consequences for the individual comprise the basic area of personality study. In general, these theories begin with the concept of the child as driven by a basic physiological nucleus of emerging and changing needs which he attempts to act out within the family constellation. The child attempts to attain direct and immediate impulse gratification, and the parents operate as socializing agents. The child's reactions to parental and other socializing agents are seen as the foundation of all later growth. The family unit and the family interrelationship are thus the basic arena for developing feelings and images about the self and modes of handling needs, and constitute the foundations of the later life style.

In adjusting to the demands of parents acting as vicars of society, certain of the child's needs are barred from expressing themselves, for example, certain sexual and hostile impulses. One of the ways in which the growing child deals with these unacceptable needs is through the mechanisms of denial or repression (discussed later in this article), often transforming the original needs into forms that are more acceptable. When such repression and transformation occur, an area of sensitivity develops in which subsequent temptation may arouse the repressed impulse and create anxiety and avoidance behavior. Needs that have undergone such repressive transformation are referred to as unconscious, in the sense that they are inadmissible and unavailable to the person. Around such repressed unconscious needs there may grow character structures that may later turn out to be the nucleus of neurotic symptoms, such as an exaggerated

concern for sexual morality in a person who has severely repressed his own sexual impulses early in development. In the sense noted here, psychodynamic theories make a continuum between normal growth and neurotic reactions, the latter being exaggerated versions of usual patterns of growth surrounded by an excess of inhibition and defense. *See NEUROSIS.*

In such theories, particularly in recent years, a more prominent place has been given to the role of ego function, the composing of conflicts in a manner that avoids the danger of neurosis. This development in theory comes in part as a reaction against the criticism that psychodynamic theories were too closely modeled upon the observation of neurotic symptoms in clinical practice with an attendant failure to consider in detail the processes of normal growth. Closer study of normal growth has indicated that there are patterns of coping that seem to be relatively independent of early difficulties suffered by the child in his first stages of socialization. Here, too, there is a reaction against the early excessively genetic-historical, family-centered approach of the psychodynamic theories. In any case, the early statement of psychodynamic theory, notably by Sigmund Freud and his followers, had in its turn tended to overreact to the nineteenth-century view of rational voluntarism and substituted an overly rigid conception of determinism (*see FREUDIANISM*). The trend since 1940 has been in the direction of reconsidering some of these views better to deal with the evidently creative aspects of growth and with the conditions that produce them. Among such studies are those that are designed to determine the kinds of family atmospheres that lead to healthy ego development, such as the development of flexible attitudes, egalitarian orientation toward others, and so on.

*Cultural approach.* The psychodynamic approach to personality has stimulated the interests of anthropologists working on early childhood socialization and has led to studies of the relationship between personality and culture. The original emphasis of these studies, as stimulated by Freud's insights and by the approach of the psychoanalytically oriented anthropologists such as R. Linton, A. Kardiner, and C. DuBois, was again upon the early development of basic personality structure. The central assumption was that certain allegedly invariant features of childhood discipline, such as mode of weaning and independence training, had the effect of creating a basic personality structure unique to a culture, and later experience and training only elaborated upon this structure. From the point of view of culture, the underlying axiom was that institutional forms, rituals, and myths were projective systems that grew out of or, in any event, matched the basic personality patterns created in the culture.

Just as in psychodynamic theory in general, in the studies of personality and culture there has been a shift away from the deterministic geneticism that placed so much emphasis upon the early ef-

fects of family life on the life pattern of the person. One finds today more emphasis upon the personality-forming effect of the roles and positions that a person fills in a society, including roles occupied during adult life, and there is at present considerable interest in the topic of adult socialization, particularly in the effects on personality produced by occupational life. Typical of such studies are the observations of N. Miller and G. E. Swanson showing the manner in which entrepreneurial families and bureaucratic families, defined in terms of the organization in which the father works, place differential emphasis upon spontaneity vs. adjustment to social requirements in the group. So, too, in the work of D. Riesman, E. Erikson, T. Parsons, and others, much more account is taken of the continuing socialization of the individual after he has entered a life of his own away from his family of origin. Erikson, for example, speaks of the stages of development of man, each representing a set of problems to be solved, pointing out that failure to solve the problems of an earlier stage makes it difficult to cope with problems of a later stage. Thus if the child does not deal successfully with the problems of coming to trust others as an infant, it is highly likely that there will be a recurrent crisis in dealing with the problems of achieving a sense of autonomy later. And given the failure of these two stages, the difficulties of realizing a sense of competence still later are compounded. But where there is a relative degree of success at each stage, the growing child or adolescent becomes increasingly free to be influenced by the myriad of social forces that may operate upon him from the society.

*Learning theory approach.* The work of learning theorists who have turned their attention to personality, such as N. E. Miller, J. Dollard, and O. Mowrer, has been strongly influenced by Freud and shares many of his basic postulates. Miller and Dollard particularly have attempted a fusion of some fundamental features of the psychodynamic approach with learning theory and with experimentation in the area of the psychology of learning (the work of C. L. Hull). Drive, response, cue, and reinforcement are conceptualized by them as the four fundamentals of learning. The processes by which behavior patterns, including neurotic symptoms, are acquired and relinquished are analyzed in terms of the details of the learning process. The development of personality consists of the learning of generalization, discrimination, and labeling responses so that the person may adjust to his environment in terms of the distinctions and equivalences that are relevant to drive reduction (*see LEARNING THEORIES*). Much of their theory of neurosis and psychotherapy bears analogous features to the earlier work of the Freudians but differs in applying experimental laboratory procedures, including the use of lower animals, for testing hypotheses about problems such as conflict and repression. Application of various models of learning (C. Hull, E. C. Tolman, and coworkers)



to the study of personality appears as a relatively current and seemingly continuous trend and is indicative of psychologists' efforts to bring such study increasingly into the realm of experimentation.

**Organismic approach.** Another important series of variations in theory which are largely compatible with the psychodynamic approach, referred to under the rubric organismic theory, has been proposed in quite different forms by K. Goldstein, A. Angyal, A. H. Maslow, G. Murphy, and W. Stern. The organismic point of view is closely related to the gestalt movement as initiated around 1910 by psychologists such as M. Wertheimer, K. Koffka, and W. Kohler, in the area of perception (see PERCEPTION). Organismic theorists share with gestalt psychologists the objection against the piecemeal analysis of behavior and they extend some of the basic principles from the area of perception to the study of the organism as a whole. Goldstein, for example, conceptualized the organism as always behaving as a unified, integrated, coherent whole, not as a series of differentiated parts. The organism is considered a single unity; what happens in a part affects the whole. Further, the individual is seen as motivated by one dominant or sovereign drive, self-actualization or self-realization, rather than by a multitude of drives. Man strives continuously to actualize his inherent potentialities by whatever channels are available to him; he does what he can do in the process of coming to terms with the world, and this single major striving gives unity, direction, and meaning to life.

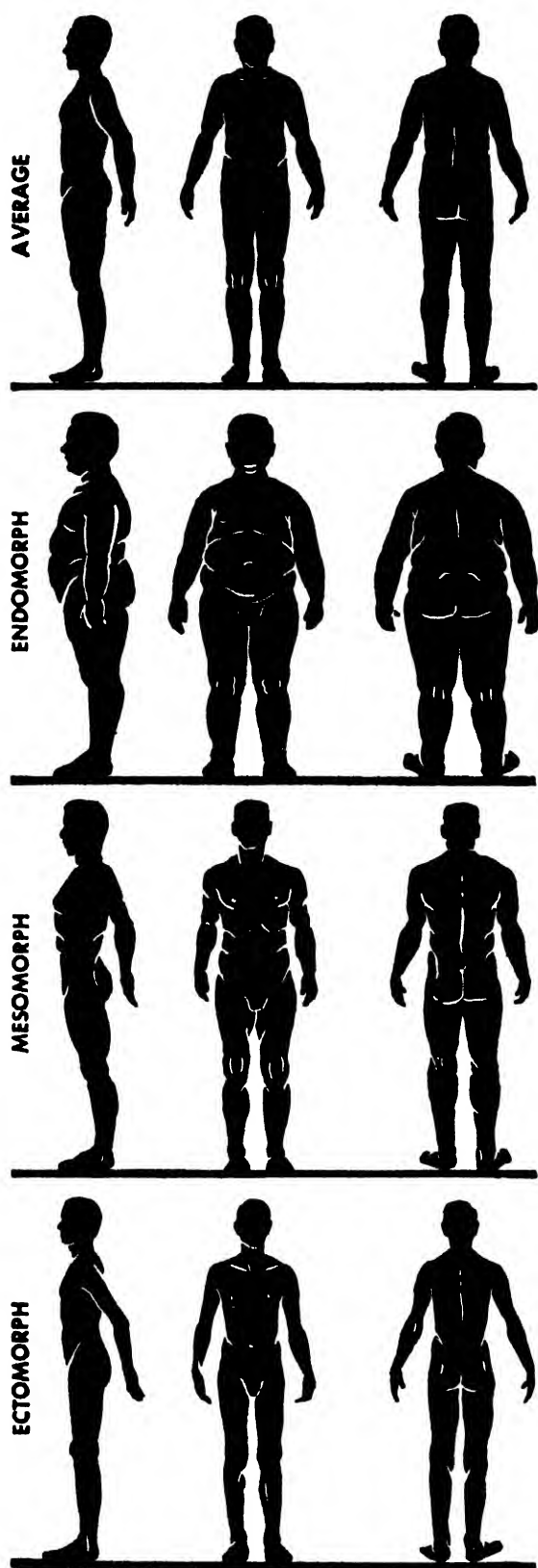
**Self-concept approach.** Self-concept theories of personality are analogous in many respects to the organismic theories (for example, C. R. Rogers, D. Snygg and A. W. Combs, and P. Lecky). The basic tenet of these theories is phenomenological in import, that is, the concept of the experienced self, a concept derived in some measure from the theories of William James and the German phenomenologists of the turn of the century. The assumption is that the striving and coping of the individual are summarized and integrated in the concept or image of the self, and that this concept or image serves as a regulatory process in guiding the growth and actualization of the person. Growth and actualization, in turn, are reflected in changes in the person's self-concept, as he moves in the direction of having an image of himself that is congruent with consistencies in his own actions and with the images that he creates in others—two aspects of insight. The person, then, strives for self-consistency, and action is hypothesized to this end. The individual achieves a mature self-concept when this concept is consistent with his experienced life pattern so that the experiences with which he must deal are not seen as alien and threatening to the self. Under these conditions, the person can use experience for further growth and self-actualization.

**Factorial trait approach.** In contrast to the above approaches, which are largely clinical and observa-

tional in method, the trait-factorial approach is based on a psychometric premise. The premise is basically taxonomic in its import, seeking a minimal set of underlying dimensions or latent structures that will describe a variety of test scores. Among the principal exponents and users of this technique of personality research are H. J. Eysenck in England and J. P. Guilford in the United States. Included in the general response tendencies confirmed by this work are such trait clusters as introversion and extraversion, general neuroticism, and so on. In recent years, an attempt has been made, particularly by Eysenck's group, to relate trait clusters or factors to certain traditional experimental variables and it has been found, for example, that introversion tends to be correlated with ease of sensory adaptation and with ability to become conditioned.

**Typological-constitutional approach.** The typological-constitutional approach is perhaps best represented in terms of two different sources. The one stems from the continuing interest in the role of bodily functioning as a determinant of temperament. The humoral approach of Hippocrates, already mentioned, continued in force throughout the medieval period. In addition, there has been a second continuing interest which is refashioned periodically as new concepts emerge. In its earliest form it was represented by the materialist doctrines of the Stoics and Epicureans who held there to be a neutral affective state from which people tended to diverge as a function of the state of their circulatory system, for example, vasodilation leading to a state of euphoria or pleasure, vasoconstriction to depression or unpleasure. There is on record a usage of this principle as a diagnostic device by the Alexandrian school of medicine, a certain caliph's distress being attributed to his relations with the members of his harem. Members of the harem were brought before him and his pulse rate recorded in order to discover the upsetting lady. In the modern period, N. Malebranche has adopted the older concept in new garb to account for the perturbations of emotion.

A similar strain of thought runs through history in the effort to relate temperamental characteristics to body type: the sluggish and indolent fat man, the long, lean, and hungry look of Cassius, and so on. Perhaps the most modern exemplification of the view is found in the work of W. H. Sheldon and collaborators. By a careful metrical analysis, they have separated body types into three components, rated on a 7-point scale, each person showing a differential loading of each component, with one of them usually dominant. The component body types are the endomorph (the round physique), the ectomorph (the narrow thin physique), and the mesomorph (the rather square athletic build). Associated with endomorphy, ectomorphy, and mesomorphy are certain temperamental traits and certain differential tendencies toward mood and mental illness. According to this theory the mesomorph tends toward overt muscular action and acting out of



Sheldon's system of body types. Extremes in ectomorphy, mesomorphy, and endomorphy are shown, as well as the average individual who has about equal proportions of all three components. (From C. T. Morgan, L. T. Alexander et al., *Introduction to Psychology*, McGraw-Hill, 1956)

his problems (somatotonia), while the ectomorph tends toward formal and theoretical formulations and intellectualization of problems (cerebrotonia). The endomorph tends to be geared more toward feeling and is relatively passive with respect both to action and cognitive activity (viscertainia). Suggestive correlational results have been obtained by this method, perhaps indicating that certain constitutional origins of personality functioning have been overlooked by the psychodynamic approach.

Another approach to the typological study of personality comes out of the German Romantic tradition, representing a development of the ideal types approach of such thinkers as W. Dilthey and the antiscientific proponents of the *Geisteswissenschaft*, the science of mind. A typical approach is that of E. Spranger, whose typological analysis of value orientations has provided the basis for one of the most widely used and productive personality tests of the present time, the Allport-Vernon-Lindzey Study of Values. Spranger's six ideal value types were the social, religious, economic, theoretical, esthetic, and political, and the Study of Values attempts to assign a set of possible points among them, on the basis of a choice of multiple responses to questions relating to preference for various activities and rewards. To a certain degree, C. G. Jung also represents this approach in his formulation of types that are polar opposites: introversion-extraversion, sensation-intuition, thinking-feeling. Jung introduces a provocative but largely untested notion of complementary functioning that is rather different from most theories: that the exercise of one extreme on a continuum sets up tensions for moving in the direction of the other extreme so that the extravert is likely to become more introverted as he goes on in life.

**Personality assessment.** The general purpose in individual personality assessment is to determine as accurately and efficiently as possible what a particular person is like with respect to some criteria. The theoretical assumptions made about the nature of personality determine both the framework in which this general question is cast and the operations selected in seeking answers. For example, if unconscious aspects of personality are considered the vital determinants of behavior, the variables and methods selected for personality study and assessment will focus on these. Or if the emphasis is on self-concept, the assessment emphasis is likely to be on phenomenological self-description. The specific purposes of the assessment, the decisions which must be made on the basis of the assessment procedure (for example, selection for executive training, for group psychotherapy, for admission to or release from a mental hospital), further constrain the procedures employed. Thus, assessment procedures are used both for research and applied purposes depending upon the choice of criterion. In the former the aim is to discover valid and generalizable relationships between various aspects of personality functioning and other factors



(for example, social class, culture, group interaction, and so on). In the latter, the aim is to reach specific decisions about individuals in industrial, military, educational, and clinical settings, and the methods tend to be as varied as the settings.

**Broad personality procedure.** In a broad personality assessment procedure, as typically used in clinical settings, the attempt is made to gain information about all of the three kinds of consistencies in individual behavior previously discussed. The individual's basic goals and needs, his characteristic style of attaining goals, and his dominant affective states are all sought through the clinician's interpretation of the individual's responses and their interrelationships, as elicited by various assessment techniques. These techniques, or tests, in which the task stimulus is structured ambiguously so as to allow maximal freedom of response (for example, a Rorschach inkblot card is shown and the person is asked what it might look like, what it might be) are termed projective. This is in contrast to those clearly structured task stimuli ("objective" tests) in which the number of possible appropriate responses is much more limited by the instructions. Projective techniques, such as the Rorschach, Thematic Apperception Test, drawing, or word association, seek to elicit a picture of the consistent, enduring aspects of the individual's inner life (fantasy, wishes, dreads, and so on) which he himself cannot verbalize or portray directly because of both social pressure within the testing situation and unconscious personal distortion and repression. The assumption underlying this indirect or projective procedure is that psychodynamic patterns are not directly accessible to the person being tested so that it is necessary to elicit behavior that will permit a reconstruction of his internal dynamics by the diagnosing clinician. The price of this indirect technique that places such heavy emphasis upon the inferences drawn by the clinician is that projective testing tends to yield relatively low interobserver reliabilities and this in turn makes the task of validating such procedures very difficult.

The task of validation is rendered the more difficult by the fact that a criterion for validation is difficult to establish within the psychodynamic framework by virtue of the emphasis on inner dynamics rather than on external, clearly specifiable performance. There is continuing debate in the field between those who propose a clinical criterion for such assessment, in which the clinicians' experience and judgment provide the basis for both inference and validation, and those who adhere to the actuarial approach which proposes that responses be rigorously categorized and checked out against an objective criterion that is consensually accepted.

**Standardized test use.** Another approach to personality assessment is through the use of standardized tests such as the Minnesota Multiphasic Personality Inventory, the Guilford GATN procedure, and R. B. Cattell's 16 P-F Test. This ap-

proach is closely linked with the trait-factorial approach to personality assessment. The procedure employed is to construct distributively adequate test items first on an intuitive basis, then to reduce the set of items by a formal or informal factorial method that gets at independent latent structures underlying responses to these items, then to select by techniques of item analysis those questions that represent the best expression of the various factors or clusters. The various trait clusters that can be tested by this method are many. Reliability is assured by this method, although the same problems of validity exist here as in the actuarial approach to projective tests, that is, finding an agreed-upon performance measure against which to pit response scores. Intelligence tests and their component factors represent but one example of this approach to assessment which, because of the pressures of practical application, tend to be seen erroneously as a different approach to assessment. The principal criticisms leveled against this approach are that it is lacking in a theoretical framework predictive of the psychodynamic aspects of personality, and that it is empirically too taxonomic in spirit and insufficiently concerned with explanation.

**Rating and ranking methods.** Rating and ranking methods are closely akin to the psychometric testing approach. Here a set of judges is armed with a criterion and asked to rank or rate a group of testees on the basis of some products produced by these testees (for example, opinions, attitudes, and expressive movements) in terms of the criterion. Reliability is determined by intrajudge and interjudge consistency, although there are special problems of errors in judgment such as the halo effect to be controlled, and the problem of validity is as great here as in the psychometric testing method. Specialized versions of this procedure, such as sociometric techniques and the Q-sort, that have been devised in recent years increase the discriminating power of ranking and rating methods.

**Naturalistic approach.** The term psychometric as it has been used in the preceding paragraphs, in contrast to the projective techniques, refers to a set of assumptions having to do with the distribution of individual differences in human performance, the assumptions of normality of distribution, of the predictable distribution of error of measurement, and of various notions concerning the scalar properties of responses to test items. The concept is thus inherently actuarial or distributional in nature, placing the individual performance in relation to an array of performances made by others rather than treating it in terms of the economy of the individual's psychodynamics. It is in terms of this distinction that G. Allport has differentiated the idiographic (individual) approach to the individual versus the nomothetic (universal) approach to generalized human functioning in which the individual represents a position or a distribution.

The naturalistic or field approach to personality assessment is well represented by the daily activity

of the field anthropologist or the diagnosing physician. It consists in drawing inferences from observable consistencies in behavior in lifelike situations. By the interposition either of formalized projective tests or of psychometric test devices, it can be converted into either of these approaches. It is the *Urmethode* of personality assessment.

**Physiological techniques.** Physiological techniques of personality assessment increasingly coming into use are mostly concerned with obtaining measures of autonomic reactivity under conditions of emotional arousal and stress. These methods are divisible into three classes: those that record electrical activity in the cortex (by use of the electroencephalograph), seeking anomalies in brain functioning; those that use a variety of measures having to do with circulation, respiration, muscular tension, skin responses, and blood chemistry; and those related to the physique of the person being tested. The second of these approaches is designed to obtain measures of motility of the autonomic nervous system and endocrinal responsiveness. This work has been strongly influenced by the theories of W. B. Cannon and, more recently, by the work of H. Selye. The emphasis of the third approach has been discussed previously in connection with the constitutional-typological theories of personality. See ELECTROENCEPHALOGRAPHY.

**Personality research.** The great bulk of personality theory is based upon clinical observation of patients in mental clinics and hospitals and child guidance clinics. Over the 25 years since 1935, although such observation has become increasingly systematic and codified, it still falls considerably short of the canons of rigor that one would demand of experimental research. This is partly in the nature of the material to be dealt with since human growth does not lend itself readily to laboratory experiment save in certain indirect ways which will be referred to shortly. The greater part of the research on which theories of personality are based can be called reconstructive in its intent, an effort on the basis of necessarily limited clinical observation to reconstruct the origins and dynamics of certain consistencies in functioning. Because of the technical problems involved, there are relatively few comprehensive longitudinal studies of personality which have followed the same individuals from early childhood into adulthood. The principal studies of this type are those being carried out by the Fels Institute and the Institute of Child Welfare of the University of California. As for other types of personality research, their approaches have already been described in connection with the discussion of personality assessment.

Another approach to personality research is through the study of those regulatory mechanisms that are necessary for adequate functioning—learning, perception, motivation, emotion, and the like. Reference has already been made to the work linking learning processes and personality processes.

Similar work has been done in connection with the contribution of other forms of functioning to growth and maintenance. Included in this work is

the study of perceptual and cognitive processes and of the manner in which these enter into the psychodynamics of the person as instruments of defense and coping. Thus H. A. Witkin and his co-workers have distinguished between people who are dependent principally upon the external visual field and those who depend upon internal bodily cues in maintaining the upright perceptually, and they show the differences in personality that are related to these orientations, the former being generally more dependent than the latter. H. Werner has shown the manner in which sensoritonic sensitivity operates in determining growth. G. S. Klein and his co-workers have shown the continuity between modes of regulation in perceptual organization and in dealing with internal drives. J. S. Bruner and his co-workers have attempted to show the manner in which perceptual selectivity serves the motivational requirements of organisms as well as their reality demands. In sum, this work has attempted since 1950 to establish a continuity between general psychological theory and personality theory by emphasizing the instrumental properties of thinking, perceiving, judging, and the like. This work has, on the whole, been closely linked with psychodynamic approaches to personality.

A similar account may be given of experimental work on specific motives. Stemming from H. A. Murray's original taxonomy of needs, there has been considerable study of the achievement motive (*n ach*) by D. C. McClelland and his collaborators, *n ach* being measured by the controlled use of questionnaires and projective techniques. This work has revealed that achievement striving is related to cultural milieu, to early training, and to the characteristic expression of anxiety in the person, notably fear of failure and desire for success. Similar work is being done with dependency needs, affiliative needs, and need for power. A closely related line of work is illustrated by the researches of E. Fromm-Reichmann on the authoritarian personality and the kinds of mechanisms that support such an orientation toward the social world, namely, intolerance of ambiguity as a cognitive and motivational characteristic.

The field of personality research is growing at such a rate that it is difficult to foretell what shape it will take in the future. Certainly there is a tendency for the different approaches to assessment to be joined together in more comprehensive programs of assessment (as, for example, at the California Institute for Personality Assessment). It is clear, too, that there is an increasing tendency for theories of personality to be less *ad hoc* and more closely linked to principles of general psychology established in more controlled experimentation (for example, objective studies of psychoanalytic mechanisms on the psychodynamic side, the work of G. Kelly on the side of transformation of subjective experience by cognitive processes, by the work of J. Piaget on child development, and by the work of N. E. Miller and of F. A. Beach on basic drive mechanisms). Finally, it seems evident that, rather

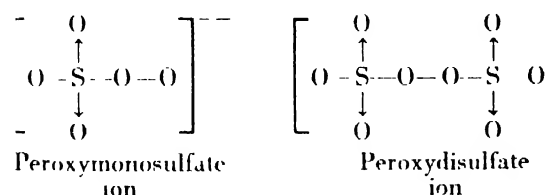
than being a special topic in psychology, the psychology of personality is rapidly becoming a central focus of general psychology concerned with overall regulatory processes that determine more segmental forms of functioning. See ABNORMAL BEHAVIOR; BEHAVIOR AND HEREDITY; BEHAVIOR, ONTOGENY OF; PSYCHOLOGY, PHYSIOLOGICAL AND EXPERIMENTAL; REPRODUCTIVE BEHAVIOR.

[J.S.B.; W.M.S.]

**Bibliography:** G. W. Allport, *Personality: A Psychological Interpretation*, 1937; C. S. Hall and G. Lindzey, *Theories of Personality*, 1957; J. McV. Hunt (ed.), *Personality and the Behavior Disorders*, 2 vols., 1944; R. L. Munroe, *Schools of Psychoanalytic Thought*, 1955; G. Murphy, *Personality: A Biosocial Approach to Origins and Structure*, 1947.

## Persulfate

A group of compounds more correctly known as peroxydisulfates. Persulfates are salts of the two peroxy acids of sulfur, peroxy monosulfuric acid ( $\text{H}_2\text{SO}_5$ ) and peroxydisulfuric acid ( $\text{H}_2\text{S}_2\text{O}_8$ ). All these compounds contain sulfur in an oxidation state of 6+, the same as in sulfates. The unusual thing about their structure is the presence of the peroxy group,  $-\text{O}-\text{O}-$ , as follows:



Ammonium peroxydisulfate is produced by the electrolytic oxidation of ammonium bisulfate solution. It is used as an oxidizing agent and a bleach, and is converted to other salts, such as potassium peroxydisulfate which is used as a polymerization promoter. See PEROXIDE; SULFUR. [E.E.WR.]

## Perthite

A parallel-to-subparallel intergrowth of potassium and sodium feldspar. With increasing size of the intergrowths, cryptoperthites, micropertthites, and perthites may be distinguished. In micropertthites and perthites the potassium feldspar is usually present as normal orthoclase to microcline, and the sodium feldspar exhibits its most ordered form as albite. In cryptoperthites the potassium feldspar is usually present as sanidine to normal orthoclase, and the sodium feldspar is present as small domains approximating analbite or albite. Connected with the small domain size in cryptoperthites a beautiful blue-to-whitish luster may be developed (moonstone). See FELDSPAR; GEM; see also ALBITE; MICROCLINE; ORTHOCLASE. [F.L.A.]

## Perturbation (astronomy)

Departure of a celestial body from the trajectory it would follow if moving only under the action of a single central force. Perturbations may be caused

by either gravitational or nongravitational forces.

**Corrections to elliptic orbits.** In the solar system, orbits of planets may be adequately represented by mean elliptical elements to which are added small corrections due to the mutual planetary attractions. Although such motion is referred to as disturbed, it is as much a consequence of the law of gravitation as is undisturbed elliptic motion, although it is much more complex.

Another method of representing perturbed motion is to augment the position derived from the mean ellipse by the actual displacements in the coordinates due to the disturbing forces. These perturbations of the elements, and perturbations of the coordinates, are represented by infinite series; usually many terms are required to represent the disturbed motion accurately. These analytical expressions are referred to as general perturbations and, with their associated mean elements, form a general theory of the motion. In some instances, such as the orbit of an outer satellite of Jupiter or the motion of a comet moving with nearly parabolic velocity, the analytical expressions representing the perturbing function become so involved (mainly because of lack of convergence of the Fourier series) that general perturbations are not attempted. Instead, the perturbed positions are computed from a step-by-step numerical integration of the equations of motion; this is known as the method of special perturbations.

**Long- and short-term disturbances.** Planetary orbits are subject to two classes of disturbances, secular, or long-term, perturbations; and the periodic, or relatively short-term, perturbations. Secular perturbations, so called because they are either progressive or have excessively long periods, arise because of the relative orientation of the orbits in space. They cause slow oscillatory changes of eccentricities and inclinations about their mean values with accompanying changes in the motions of the nodes and perihelia. The periods of time involved in these oscillations may extend from 50,000 to 2,000,000 years. Periods and major axes of orbits are not affected by secular changes. For the orbit of the Earth, the present inclination to the invariable plane is  $1^\circ 35'$ . This will diminish to a minimum of  $47'$  in approximately 20,000 years. The eccentricity, presently 0.017, is diminishing also and will reach a minimum of 0.003 in about 24,000 years.

Periodic perturbations arise from the relative positions of the planets in their orbits. When the disturbed and disturbing planets are aligned on the same side of the Sun, the perturbation reaches a maximum, and reduces to minimum when alignment is reached on opposite sides of the Sun. The size of a periodic perturbation is a function of the mass of the disturbing body and of the length of time the two planets remain near the point of closest approach. Periodic perturbations continually shift a planet away from the position it would occupy in undisturbed motion, moving it above or below the orbital plane, nearer to or farther from the Sun, and forward or backward in the orbit.

**Commensurable motions.** If the mean motion of the disturbed planet were exactly a submultiple, say  $\frac{1}{2}$ , of the mean motion of the disturbing planet, the maximum perturbation produced by their close approach would always occur in the same part of the disturbed orbit. The displacement in position of the disturbed planet would increase with each coincidence until the character of the orbit became modified to the point where exact commensurability of the mean motions would cease to exist.

Because the solar system is middle-aged, cosmically speaking, few examples of commensurability of mean motions exist today. None is found in the motions of the major planets. Cases of near commensurability exist which give rise to long-period periodic terms of large amplitude. As an example, the periods of Jupiter and Saturn are nearly in the ratio of 2:5. Thus, after nearly 5 revolutions of Jupiter, the two planets return to approximately the same juxtaposition. Their line of coincidence, however, sweeps slowly around Jupiter's orbit, completing a circuit in about 850 years and thus producing a perturbation of this period.

Among the four inner planets, the periodic perturbations are small, amounting in orbital longitude at most to 225' for Mercury, 45' for Venus, 1' for Earth, and 2' for Mars. Periodic perturbations of the outer planets are larger, reaching in the case of the long-period terms to 30' for Jupiter, 70' for Saturn, 60' for Uranus, and 35' for Neptune.

Because the amplitude of a periodic perturbation depends on the mass of the disturbing planet, observational measurement of this amplitude affords a method of determining the disturbing mass. For the planets Mercury, Venus, and Pluto, which do not have satellites, this is the only method of determining the mass. As a consequence of the mutual perturbations of the planets, the distance of a planet from the Sun is, on the average, decreased by the action of planets closer to the Sun, and increased by planets farther from the Sun; this mean effect represents a perturbation of the radius vector with a constant value.

The orbits of the minor planets are affected in varying degree by the attractions of the major planets. Those orbits passing close to Jupiter suffer large perturbations which, if the mean motions were commensurable with that of Jupiter, would be augmented at each close approach until the trajectories were sufficiently altered to reduce the commensurability. In the over-all distribution of mean motions of the minor planets there are noticeable gaps at the points where the period would be nearly an exact submultiple ( $\frac{1}{2}$ ,  $\frac{1}{3}$ ,  $\frac{1}{4}$ , . . .) of the period of Jupiter. In cases of near commensurability, observational determination of the amplitude of the long-period perturbation affords a method for measuring the mass of Jupiter. A small group of minor planets, called the Trojan asteroids, has been so completely captured by Jupiter that they oscillate about the 60° points which form

equilateral triangles with Jupiter and the Sun. See TROJAN PLANETS.

**Effect on comets.** Planetary perturbations also affect the orbits of comets. Studies of the motion of Halley's comet indicate that the time from one perihelion passage to the next has varied by almost five years because of perturbations. Most comets approach the Sun at nearby parabolic speeds in randomly oriented orbits, but if a comet approaches close to one of the more massive major planets, the planet may so alter the trajectory that the comet pursues an elliptical orbit thereafter. A number of short-period comets whose orbits agree only in that they all pass close to Jupiter illustrate the perturbing effect of this planet on cometary orbits.

**Nongravitational causes.** Material forming the tails of comets is subject to a nongravitational type of perturbation. This rarefied matter given off by the head of the comet is forced into a trajectory away from the Sun by the pressure of solar radiation.

Associated with many of the periodic comets are swarms of smaller particles which appear as meteors upon collision with the upper atmosphere of the Earth (see METEOR). The density of these swarms is so tenuous that they cannot hold themselves together by their own gravitation, and planetary perturbations of speed and direction soon spread the components completely around the orbit. The annual meteor showers, such as the Perseids, reflect this dispersal of particles along the orbit. The effect of the Earth's attraction on a meteor trajectory depends on the relative velocity; that is, whether the Earth is overtaking the meteor or meeting it head on. Once the meteor enters the upper reaches of the Earth's atmosphere its motion is subject to a nongravitational perturbation caused by atmospheric drag. This resistance to the passage of the particle is evidenced by the trail of incandescent gas and vapor which forms until the particle is consumed or continues in its trajectory greatly decelerated.

**Perturbations of satellite orbits.** The motions of planetary satellites, natural and artificial, reflect both gravitational and nongravitational perturbations. The centrifugal force arising from the rotation of a planet causes a deformation or oblateness of figure. In such a case the central mass does not attract as if it were concentrated at its center. For a close satellite the principal perturbation arises from the attraction of this equatorial bulge. The effect of this attraction on an otherwise undisturbed satellite orbit is a gradual regression of the line of nodes on the equatorial plane and a rotation on the line of apsides. Both rotations vary with the inclination of the satellite orbit. Nearer to the primary, the tidal forces may become so great that a satellite would be literally torn to pieces. For a fluid satellite of the same density as the planet, the limit within which this disruptive perturbation occurs is about two and one-half times the radius of the planet (see SATURN).

Satellite motions are also disturbed by the direct attraction of other satellites, the Sun, and, to a lesser amount, by other planets. Observation of the orbital displacements caused by the mutual perturbations of satellites in the systems of Jupiter and Saturn makes possible the determination of the masses of these satellites. The solar attraction is significant in the orbits of the outer satellites of Jupiter and Saturn, reaching to one-ninth the planet's attraction for the eighth satellite of Jupiter. So greatly disturbed is this satellite that it is not possible to derive a general theory for its motion.

The orbit of the Moon is disturbed mainly by the Sun, with some changes in motion due to the oblateness of the Earth, the figure of the Moon, and smaller perturbations caused by the planets. The attraction of the Sun on the Moon is more than twice the Earth's attraction, but because both the Earth and Moon are free to move it is only their relative acceleration with respect to the Sun which determines the motion. This relative acceleration toward the Sun is always less than  $\frac{1}{60}$  of the acceleration of the Moon toward the Earth. The eccentricity and inclination of the Moon's orbit oscillate slowly about their mean values, while the line of apsides advances with an average period of almost 9 years and the nodes regress through one revolution in 18.6 years.

The observed motion of the lunar node and perigee affords one means of measuring the oblateness of the Earth; the lunar theory which best represents this motion incorporates the value  $\frac{1}{601}$ . The value of the flattening of the Earth adopted for the International Reference Ellipsoid, on the basis of geodetic and gravimetric measurements, is  $\frac{1}{607}$ . Recently, measures of the motion of the nodes and apsides of artificial Earth satellites have indicated a value of the oblateness of  $\frac{1}{608}$ . Lunar and solar perturbations of artificial Earth satellite orbits are minor for orbits 500 miles above the surface but grow with increasing distance from the Earth. Atmospheric drag perturbations are significant at this altitude, however, and decrease with increasing altitude. *See* CELESTIAL MECHANICS.

[R.L.D.]

*Bibliography:* J. A. Van Allen (ed.), *Scientific Uses of Earth Satellites*, 2d ed., 1958.

## Perturbation (mathematics)

A modification in the mathematical structure of a problem changing the problem from one that can be solved exactly, the unperturbed problem, to one, the perturbed problem, for which it is usually possible to obtain only an approximate solution. The methods employed for this purpose form perturbation theory. These methods attempt to express the solution of the perturbed problem in terms of the properties of the solutions of the unperturbed problem.

**Examples.** Examples of perturbation problems can be found in nearly every branch of mathematics and physics, and in astronomy. For the latter, *see* PERTURBATION (ASTRONOMY). The simplest case

occurs in ordinary algebra. Suppose that the roots of the equation  $f(x) = 0$  are known (the unperturbed problem), and that the roots of the equation  $f(x) + \epsilon g(x) = 0$  are to be found (the perturbed problem). The parameter  $\epsilon$  measures the size of the perturbation. Another set of examples occurs in linear differential equations and in particle dynamics. Possible perturbations include changes in the forces considered to be acting on the particle as well as changes in initial conditions.

Several examples occur in partial differential equations. One physical realization occurs in the theory of wave propagation where the perturbations can be changes in the index of refraction, changes in initial conditions, or changes in the nature or shape of the surfaces encountered by the waves. All of these changes can occur separately or concurrently. The first of these changes is called a volume perturbation, the second a perturbation of initial conditions, the third a perturbation of boundary conditions. Similar examples can be taken from quantum mechanics, where the volume perturbation corresponds to a change in the Hamiltonian, and perturbation of initial conditions to quantum mechanical time-dependent perturbation theory. *See* PERTURBATION (QUANTUM MECHANICS). Other partial differential equations of physics such as the Laplace equation, the diffusion equation, and the equations of hydrodynamics furnish further examples.

As a final illustration of these various types of perturbation, consider possible modifications in an equation (as well as boundary and initial conditions) describing the motion of particles such as neutrons or electrons moving through a medium which can scatter and absorb them. The equation is known as the Lorentz-Boltzmann equation and changes in it occur as a consequence of modifications of the laws of scattering and absorption, that is, because of changes in the medium.

All of these problems are linear and can therefore be cast into an equation of the form  $A\psi = \lambda\psi$ , where  $\psi$  is the unknown quantity,  $\lambda$  is a constant, and  $A$  is an operator involving among other possibilities differentiation and integration. The quantity  $\psi$  may be a scalar, vector, or more generally a matrix quantity. When solutions can be obtained for only special values of  $\lambda$ , the eigenvalues, the equation is called the eigenvalue equation, and the associated problem, the eigenvalue problem. The operator  $A$  contains the perturbation; that is,  $A$  equals  $A_0 + \epsilon A_1$ , where  $A_0$  is the unperturbed operator and  $\epsilon A_1$ , the perturbing term.

**Iteration method.** The method generally employed to obtain an approximate solution is called the iteration method. Rewrite the equation  $A\psi = \lambda\psi$  as  $(A_0 - \lambda)\psi = -\epsilon A_1\psi$ . Let the unperturbed solution be  $\phi_0$ , where  $(A_0 - \lambda_0)\phi_0 = 0$ . Then  $\psi_1$ , a first approximation to  $\psi$ , is obtained as a solution of  $(A_0 - \lambda)\psi_1 = -\epsilon A_1\phi_0$ . A second approximation is the solution of  $(A_0 - \lambda)\psi_2 = -\epsilon A_1\psi_1$ . The  $n$ th approximation is obtained in terms of the  $(n-1)$  approximation,  $\psi_{n-1}$ , from the equation

$(A_0 - \lambda)\psi_n = -\epsilon A_1 \psi_{n-1}$ . It is assumed that the properties of the unperturbed operator,  $A_0$ , are completely known so that the solution of these equations can be obtained. If the sequence  $\phi_0, \psi_1, \dots, \psi_n, \dots$ , converges, it will converge to a solution of the problem. For an eigenvalue problem, the procedure must be modified. The first approximation to  $\lambda$  is  $\lambda_0$ ; the  $n$ th approximation is  $\lambda_n$ . Then the equation determining  $\psi_n$  in terms of  $\psi_{n-1}$  is  $(A_0 - \lambda_{n-1})\psi_n = -\epsilon A_1 \psi_{n-1}$ . It is important for the practicality of this procedure that the approximation  $\lambda_n$  can be expressed in terms of the approximation  $\psi_{n-1}$  and the operators  $A_0$  and  $\epsilon A_1$ .

In a related and more familiar formulation both  $\psi$  and  $\lambda$  are expanded in a power series in  $\epsilon$ , that is,  $\psi = \phi_0 + \epsilon \phi_1 + \epsilon^2 \phi_2 + \dots$  and  $\lambda = \lambda_0 + \epsilon \lambda_1 + \epsilon^2 \lambda_2 + \dots$ . Then the equation  $A\psi = \lambda\psi$  reduces to a set of equations for  $\phi_n$ . For example, the equation for  $\phi_1$  is  $(A_0 - \lambda_0)\phi_1 = -(A_1 - \lambda_1)\phi_0$  and the equation for  $\phi_2$  is  $(A_0 - \lambda_0)\phi_2 = -(A_1 - \lambda_1)\phi_1 + \lambda_2\phi_0$ , and so on. This formulation is more complex, and often yields slower rates of convergence than the method just outlined.

The iteration method can be generalized in two respects. First, it is not necessary to use  $\phi_0$  as the zeroth approximation to  $\psi$ . If by reason of other information, a better approximation, say  $\psi_0$ , is known, the iteration sequence starts with the equation  $(A_0 - \lambda)\psi_1 = -\epsilon A_1 \psi_0$ . Second, the iteration method can be employed in the treatment of nonlinear as well as the linear problems discussed in detail here.

For the iteration method to be at all possible, it is necessary for the sequence  $\psi_0, \psi_1, \dots$  to exist and to converge. The usefulness of the method increases with increasing rate of convergence. The sequence exists only if the singularities of the perturbation are not too strong, or if the initial zero approximation is properly chosen, or both. When the sequence exists, it will converge for a range in values of the parameter  $\epsilon$ . The largest value of  $\epsilon$  is the radius of convergence. This is found to be that value of  $\epsilon$  for which the equation  $A\psi = \lambda\psi$  has at least two degenerate solutions, that is, solutions with identical values of  $\lambda$ . There are various methods of increasing the radius of convergence. For example, the general techniques of analytic continuation, such as the Euler transformation, can often be employed. A clever choice of  $\psi_0$ , the zeroth approximation, will often produce the desired effect. The variational method can generate the appropriate choice for  $\psi_0$ . A more general method was developed by I. Fredholm in which the solution of  $A\psi = \lambda\psi$  is given as the ratio of two functions, each of which can be expressed as a series in  $\epsilon$ . For a wide class of operators  $A$ , each of these series will have an infinite radius of convergence.

The eigenvalue problem can be reduced to the problem of the solution of a set of homogeneous linear simultaneous equations which is generally but not always infinite. A nontrivial solution of these equations is possible only if the determinant

of the coefficients is zero. Because the coefficients involve the eigenvalue  $\lambda$ , this condition yields an equation, the secular equation, which determines the possible values of  $\lambda$ . The determinant is known as the secular determinant. An example follows. Let the solutions of the unperturbed problem be  $\phi^{(p)}$  with eigenvalues  $\lambda^{(p)}$ ; that is,  $A_0\phi^{(p)} = \lambda^{(p)}\phi^{(p)}$ . Moreover, suppose that the set  $\phi^{(p)}$  is complete, which roughly means that an arbitrary function can be represented as a linear combination of  $\phi^{(p)}$ . Therefore let  $\psi$ , the solution of the perturbed problem, be  $C_0\phi^{(0)} + C_1\phi^{(1)} + C_2\phi^{(2)} + \dots$  where  $C_p$  are constants. By substituting this expression for  $\psi$  in the equation  $A\psi = \lambda\psi$  and employing the properties of the set  $\phi^{(p)}$  which follow from the nature of the operator  $A_0$ , one can obtain a set of equations for  $C_p$ . In a typical case these equations have the following form:

$$\begin{aligned} C_0(\lambda - \lambda^{(0)}) + C_1(\epsilon A_1)_{10} + C_2(\epsilon A_2)_{20} + \dots &= 0 \\ C_0(\epsilon A_1)_{01} + C_1(\lambda - \lambda^{(1)}) + C_2(\epsilon A_1)_{21} + \dots &= 0 \\ C_0(\epsilon A_1)_{02} + C_1(\epsilon A_1)_{12} + C_2(\lambda - \lambda^{(2)}) + \dots &= 0 \end{aligned}$$

and so on. The elements  $(\epsilon A_1)_{pq}$  are numbers which depend upon  $\phi^{(p)}$ ,  $\phi^{(q)}$  and the operator  $\epsilon A_1$ . The consequent secular equation is obtained by setting the determinant of the coefficients of  $C_p$  in this sequence of equations equal to zero. The solution of these simultaneous equations for the coefficients  $C_p$  can be obtained by the iteration method, which yields a particular representation of each of the approximations  $\psi_n$ . If there are only a finite number of  $C_p$ , the secular determinant is of finite order and reduces to a finite polynomial in  $\lambda$  so that in the finite case solutions of the secular equation can always be obtained without recourse to perturbation methods. Once the allowed values of  $\lambda$  are known, the corresponding values of  $C_p$  can be determined.

**Degenerate perturbation theory.** A special technique is required when the unperturbed problem is degenerate, that is, when there are several solutions of the equation  $A_0\phi = \lambda\phi$  for a single value of the eigenvalue  $\lambda$ . The number of such independent solutions is the order of the degeneracy. The corresponding method is designated degenerate perturbation theory. The objective of the special method adopted for this case is the determination of the appropriate linear combinations of these degenerate solutions for use as the initial approximation,  $\psi_0$ , in the iterative method. To this end, all terms in the equations for  $C_p$  are dropped when  $p$  refers to an unperturbed solution which is not one of the degenerate solutions under consideration, and only those  $C_p$  which do refer to these degenerate solutions are retained. The resulting secular equation for  $\lambda$  has a number of roots equal to the order of the degeneracy. For each root there is a corresponding set of values for  $C_p$  which determine a particular linear combination of the degenerate unperturbed solutions. Each of these combinations



can be employed as the initial approximation,  $\psi_0$ , in the iterative method. It is often the case that the determination of the possible  $\psi_0$  is sufficient for the evaluation of the major effects of the perturbation. [H.F.E.]

**Bibliography:** P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, 1953.

## Perturbation (quantum mechanics)

Perturbation techniques are used in quantum mechanics, as in astronomy, classical mechanics, and other fields, to obtain approximate solutions to problems. The vast majority of quantum mechanical problems cannot be solved exactly. The techniques used are numerous and mathematically intricate; only a few are described here. For background material see QUANTUM MECHANICS; QUANTUM THEORY, NONRELATIVISTIC. Although the discussion here is confined to the domain of nonrelativistic quantum theory, where quantum perturbation theory is concerned almost exclusively with solutions to Schrödinger's equation, there have been important applications to quantum electrodynamics and quantum field theory. See QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY.

In many problems, Schrödinger's equation has the form  $H\psi_n \equiv (H_0 + \lambda V)\psi_n = E_n\psi_n$ ; the desired eigenvalue  $E_n(\lambda)$  and eigenfunction  $\psi_n(\lambda)$  are in the discrete spectrum; and the corresponding  $\epsilon_n = E_n(0)$  and  $u_n = \psi_n(0)$  solving the unperturbed equation  $H_0 u_n = \epsilon_n u_n$  are known exactly. One assumes that  $E_n$  and  $\psi_n$  can be expanded as power series in  $\lambda$ , substitutes these series (whose coefficients are still undetermined) into the Schrödinger equation, and successively equates terms proportional to the same power of  $\lambda$ , starting with  $\lambda^0$ . This procedure successively determines the unknown coefficients in terms of  $\epsilon$  and  $u$  and appears to converge best when

$$\lambda V'_{nm}(\epsilon_n - \epsilon_m)^{-1} \ll 1$$

where  $V'_{nm} \equiv \int u_n^* V' u_m$  are the matrix elements of  $V'$  in the unperturbed representation. The function  $u_n$  must be and can be chosen so that  $V'_{nm} = 0$  whenever  $\epsilon_n = \epsilon_m$ , that is, the perturbation must not "mix" originally degenerate states; this requirement selects the set  $u_n = \lim \psi_n(\lambda)$  as  $\lambda \rightarrow 0$  from the many possible degenerate  $u_n$ . This perturbation method is employed, for example, to estimate the level splitting in the Zeeman effect; as another illustration of its utility, one infers that associating atomic energy levels with unique configurations of single-particle states is most likely to be a good approximation, that is, is least likely to neglect serious configuration mixing, when the single-particle levels are widely spaced.

A similar expansion in powers of  $\lambda$  is used in collision problems. Here  $\psi(E) = u(E) + \lambda v(E, \lambda)$  is sought, where the scattered wave  $\lambda v$  satisfies an assigned (usually the outgoing) boundary condition; the energy  $E$  is known and lies in the continuum where the level spacing is zero, so that the previous procedure requires modification.

The term proportional to  $\lambda$  in the series expansion of  $\lambda v$  yields the first Born approximation to the scattering cross section; the term proportional to  $\lambda^2$  yields the second Born approximation, etc. The sequence of Born approximations appears to converge best when  $\lambda V'/E \ll 1$ , that is, when the incident kinetic energy is high; exceptions to this rule do occur, however, especially in many-particle collisions. See SCATTERING EXPERIMENTS, ATOMIC AND MOLECULAR; SCATTERING EXPERIMENTS, NUCLEAR.

The WKB (Wentzel-Kramers-Brillouin) method substitutes  $\psi = A \exp(2\pi i W/\hbar)$  into Schrödinger's equation, and assumes  $A$ ,  $W$  are expandable in powers of  $\hbar$ . In zeroth approximation, as  $\hbar \rightarrow 0$ , the phase  $W$  satisfies the Hamilton-Jacobi equation, demonstrating the connection between classical and quantum mechanics (see HAMILTON-JACOBI THEORY). The practical utility of the next (WKB) approximation, giving the quantal corrections to the classical limit  $\hbar \rightarrow 0$ , is almost exclusively confined to one-dimensional problems; the WKB approximation is best when  $d|\lambda|/dx \ll 1$ , where  $|\lambda(x)| \equiv \hbar/|p| = \hbar[2m|E - V|]^{-1/2}$  is the magnitude of the local de Broglie wavelength. For continuum eigenfunctions the WKB approximation yields estimates of the penetrability of a barrier and of the phase shifts in scattering problems. For bound state eigenfunctions the WKB approximation yields a corrected version of the pre-Schrödinger quantization rule  $\oint p dx = nh$ .

The foregoing methods pertain to time-independent problems. When the perturbation is a time-dependent interaction  $V'(t)$ , one usually wishes to compute the transitions which  $V'$  induces from initial stationary states  $u_i$  of  $H_0$  to final stationary states  $u_f$ . The "sudden approximation" is adapted to potentials  $V'(t)$  which change during a brief interval  $\Delta t$ , and appears to be valid when  $\omega_{fi}\Delta t \ll 1$ , where  $\omega_{fi} = \hbar^{-1}(E_f - E_i)$  are the circular frequencies associated with the possible transitions  $i \rightarrow f$ . This approximation is useful, for example, to determine the probability that radioactive decay of a  $\text{H}^1$  atom will yield a doubly ionized  $\text{He}^2$  ion. For slowly changing potentials ( $\omega_{fi}\Delta t \gg 1$ ), the "adiabatic approximation," which supposes that discrete quantum numbers characterizing the initial wave function remain constant as the potential changes, often is useful. See PERTURBATION (ASTRONOMY); PERTURBATION (MATHEMATICS). [E.G.]

**Bibliography:** See QUANTUM MECHANICS; QUANTUM THEORY, NONRELATIVISTIC.

## Pesticide

A material used to control insect, disease, and weed pests which cause significant losses in agriculture. Crop losses in the United States caused by insects, disease, and weeds have been estimated to total more than \$12,000,000,000 annually. This loss represents approximately one-third the total annual value of agricultural production in the United States. See AGRICULTURAL CHEMISTRY; AGRICUL-

TURE; FUMIGANT; FUNGISTAT AND FUNGICIDE; HERBICIDE; INSECTICIDE; MITICIDE; NEMATOCIDE; RODENTICIDE. [C.A.H.; O.D.]

## **Petrifaction**

One of the most remarkable mechanisms by which the remains of extinct organisms are preserved in the fossil record is the process of petrification. In petrifications, though chiefly in the case of plants rather than animals, there is retained relatively undeformed the original shape and topography of the tissues and occasionally even minute cytological details.

**Formation.** The term petrification was adopted as a scientific term before knowledge existed of the geochemical mechanism or processes involved. It was formerly widely believed that in the formation of a petrification the organic matter of the organism or tissue involved was replaced molecule by molecule with mineral material entering in solution in percolating ground water. It is now evident that what actually happens is that the mineral fills cell lumina and the intermicellar interstices of cell walls with insoluble salts depositing from solution. Petrification is hence a form of mineral emplacement or embedding by which the organic residues are filled with solid substance which infiltrate in solution. The most common substances involved in petrifications are silica,  $\text{SiO}_2$ , and calcium carbonate,  $\text{CaCO}_3$  (calcite). Occasionally phosphate minerals, pyrite, and hematite and other less common minerals comprise all or part of the petrification matrix. The most perfect preservation of original structure is found in siliceous petrifications. The clear, transparent, or microcrystalline silica renders excellent optical properties to thin sections of such fossils and makes possible the use of transmitted light in microscopic study, in much the same manner as with recent biological material.

**Calcified types.** An unusual type of calcareous petrification known as coal balls occurs in Carboniferous coal seams of parts of Europe and North America. They comprise nodular, usually spheroidal or ovoidal masses of relatively uncompressed plant tissues completely permeated with calcite or dolomite. They represent irregularly spaced and localized areas of mineral precipitation with resulting petrification of the coal-forming plant debris. Mineralization occurred while the coal was still in the peat stage. After mineralization the plant parts so infiltrated failed to compress so that their structure and cellular organization are preserved. Much of what we know of the internal organization and anatomy of ancient plants and evolution of their organs, both vegetative and reproductive, is derived from petrifications, which occur throughout the geologic record from Precambrian to recent. Coal balls, though limited to the Carboniferous, have provided an unusually comprehensive body of knowledge on the morphology and anatomy of the rich and varied vegetation of this unit of geologic time. See COAL BALLS.

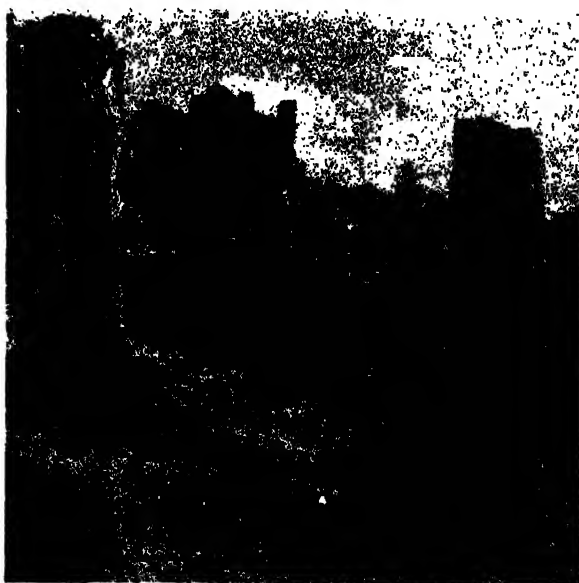
**Silicified types.** Despite the widespread abundance of petrifications (chiefly of plants but also

of the hard parts of animals) throughout the geologic column, there is very little known of the geochemical mechanisms which induce their formation. The problem is particularly baffling in the case of silicification, in many examples of which entire trunks of trees may become completely mineralized with no visible evidence of the sequence of changes following sedimentation. The fact that silicified plant parts often show virtually no physical distortion or compression indicates that the process occurs relatively rapidly. On the other hand, the amount of silica in ground water ( $\pm 70$ –100 parts per million) is so small that the process must proceed with extreme efficiency. It is probable that decay processes must have proceeded to at least the early stages since the histology of silicified (also calcified) plant tissues exhibits varying stages of degradative alteration to the extremes of nearly total loss of structure. The percentage by weight of organic matter retained in silicified wood may range from a few hundredths of 1% to more than 15%. Ordinarily it is only a few per cent. See FOSSIL; PALEOBOTANY; PETRIFIED FORESTS. [E.S.B.]

## **Petrified forests**

Exposures containing appreciable numbers of petrified tree trunks, either standing upright or lying prostrate in the enclosing sedimentary rocks; sometimes called "fossil forests." The best known examples are the widely known Petrified Forest of Arizona, and the fossil forests near Cairo, Egypt; near Calistoga, California; near Vantage Bridge, Washington; and in Yellowstone Park, Wyoming.

The Petrified Forest of Arizona is made up of hundreds of prostrate silicified tree trunks, logs, and stumps lying scattered at random like an ancient "log drive." These are of Triassic age, roughly 175,000,000 years old, and occur in a portion of the "Painted Desert," which owes its brilliant hues to



Upright petrified trees, Specimen Ridge, Yellowstone National Park, Wyoming.



the varicolored layers of the Chinle formation. The wood is mainly agatized or changed to chalcedony and shows an unusually varied and beautiful coloration of reds, browns, yellows, and purples. The majority of the petrified trees are Conifers (*Araucarioxylon*), distantly related to the araucarian "pines" of South America and Australia. One huge log over 100 ft long has been left by erosion across a ravine about 40 ft wide, forming a natural span known as Agate Bridge. Here and there are exposures of shale beds containing many impressions of leaves and seeds representing a humid, subtropical forest bordering the streams of a lowland savanna.

An even more extensive fossil forest lies in the northeastern portion of Yellowstone National Park, Wyoming. Here the majority of the petrified tree trunks are found still standing upright in positions of original growth in the enclosing medium of volcanic tuffs and breccia. Even more unusual is the occurrence here of not merely a single fossil forest, but a vertical succession of over 20 buried forests—one above the other—in a thickness of over 2000 ft of volcanic debris. The fossilized trees and the impressions of leaves, twigs, needles, and cones in associated ash layers indicate a forest of over 100 species of warm temperate to subtropical trees typical of a humid, lowland environment. See PALEOBOTANY; see also VEGETATION ZONES. [E.D.]

## Petrochemical

One of a large number of substantially pure chemical substances produced commercially from petroleum or natural gas. Ordinarily, the term does not include hydrocarbon fuels and lubricants, nor chemicals produced by others than the processor handling the petroleum raw material. See PETROLEUM PRODUCTS. Organic compounds comprise the great bulk, as well as number, of petrochemicals, but several inorganic compounds (ammonia, carbon black, sulfur, and hydrogen peroxide) also are produced in large amounts.

Thus, petrochemicals are not to be regarded as a particular type or class of chemical, since all of them have been, and many still are, made from other raw materials; for example, benzene and acetylene from coal, glycerol from fats, ethyl alcohol from agricultural crops, and sulfur from deposits of the element or from metal ores.

Some chemicals once made from other raw materials are now made entirely, or almost entirely, from petroleum or natural gas. Examples are acetone, originally derived from wood distillation and later by fermentation of agricultural products; ethyl chloride, originally made from ethyl alcohol produced by fermentation; and butadiene, made from ethyl alcohol during World War I and, as a stopgap measure, about half from ethyl alcohol during World War II, but since 1946 derived almost completely from petroleum.

Petrochemicals include also many products never before known except in laboratory amounts, such as isopropyl alcohol, ethylene oxide, glycol ethers, allyl chloride, allyl alcohol, epichlorohydrin, methyl isobutyl ketone, and acrolein.

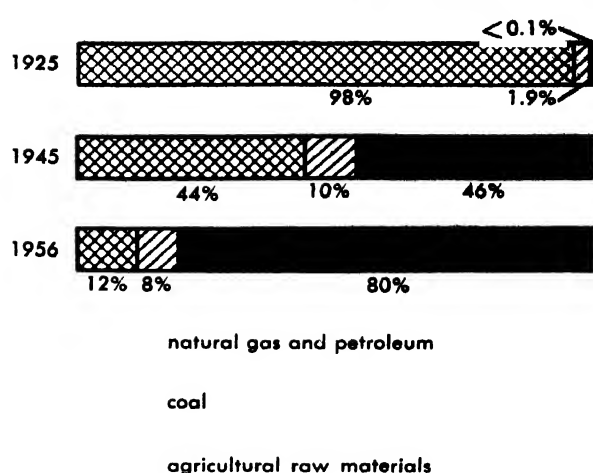


Fig. 1. Sources of organic chemicals in the United States.

**Growth and present size.** The over-all trend to petroleum and natural gas as the predominant source of organic chemicals in the United States is shown in Fig. 1. Two of the inorganic petrochemicals (ammonia and carbon black) also dominate their respective fields.

Over 40,000,000,000 lb of petrochemicals are produced annually, about 27% of the total chemical production in the United States (organic and inorganic), and from one-third to over one-half of its total dollar value, depending upon the basis used. Petrochemicals manufactured in the United States make up over 85% of all aliphatics, 10% of all inorganics, and 54% of all aromatics. About 3% of all crude oil used in the United States is used as petrochemical feed-stock.

There were no petrochemicals before 1920 aside from carbon black, which has been made in the United States from natural gas since 1872. The growth of petrochemicals since 1920 (Fig. 2) has been most spectacular in the field of aliphatics, which account for well over half of the present total volume of petrochemicals and by far the largest number of the nearly 3000 petrochemical-derived compounds now produced. This growth is due mainly to the following five factors:

1. The abundance and relatively low cost of crude oil and natural gas.

2. Advances made in the technology of petroleum refining, spurred especially by the demand for motor gasoline and aviation gasoline. These include more efficient fractional distillation and other separation processes, particularly for the lower-boiling constituents and the development of conversion processes to increase gasoline quantity and quality, including thermal and catalytic cracking, hydrogenation, dehydrogenation, polymerization, isomerization, and alkylation. These separation and conversion processes have made available as by-products, or have been readily adapted to the intentional production of, large quantities of individual hydrocarbons or simple mixtures quite suitable as chemical raw materials.

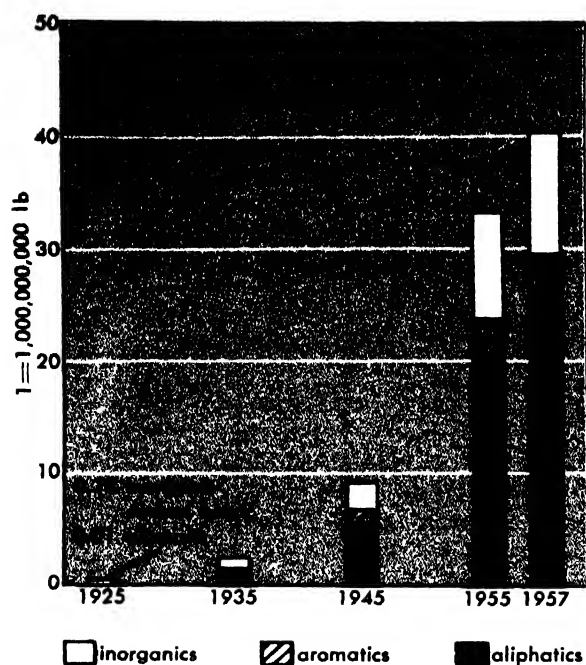


Fig. 2. Growth of petrochemicals production in the United States.

3. A demand for chemicals that in many cases could not be supplied in sufficient amounts, or with sufficient assurance of a steady supply and price structure, from other sources—coal, wood, or agricultural products.

4. Concentration of end uses in the relatively young and more rapidly expanding sectors of the economy; for example, automotive and aviation chemical products such as antifreeze and anti-knock compounds, and brake fluids; synthetic fibers; plastics and resins; and protective coatings.

5. Research-mindedness, leading to products not before known commercially (some of which have found extensive use), also to lower-cost processes for established products, often greatly expanding their application. A special characteristic of petrochemical research has been the development of a series of related derivatives from the primary petrochemicals in order to establish the widest possible market.

**Raw materials and products.** The major operations of the petrochemical industry are summarized below for each of the principal raw materials. In addition to the products shown, large numbers are made in smaller amounts.

Natural gas varies widely in composition but consists predominantly of methane, with successively smaller amounts of higher paraffin hydrocarbons. Hydrogen sulfide, carbon dioxide, and nitrogen are sometimes present. The principal petrochemicals derived from natural gas or its methane content are shown in Fig. 3 and Table 1.

**Inorganic petrochemicals.** Ammonia is by far the largest volume petrochemical, and natural gas methane is by far the largest source of the hydrogen, although hydrogen from refinery sources such

as catalytic reforming is growing in use. A large proportion is converted into ammonium nitrate and other ammonium salts and into urea. The largest use of these, as of ammonia itself, is as fertilizer.

Carbon black is made almost entirely by partial combustion with insufficient air supply, using natural gas in the old channel process and either natural gas or highly aromatic petroleum oil in the newer furnace process. The latter process (accounting for about 75% of the carbon black production) gives a product suitable for use in synthetic rubber. See CARBON BLACK.

**Aliphatic organic compounds.** Methane is the chief source of methanol and hydrogen cyanide, and an important source of chloromethanes and acetylene.

Paraffin hydrocarbon raw materials heavier than methane are mainly ethane, propane, butane, and pentane. The first two, especially ethane, are cracked thermally in large quantities to produce ethylene (and propylene), as discussed below. Propane is also nitrated commercially to a mixture of nitroparaffins, which have found small-scale use as solvents. Propane and butane are oxidized directly in fairly large amounts to formaldehyde and acetaldehyde, with minor amounts of other oxygenated hydrocarbons. Small amounts of pentanes are chlorinated; the chloropentanes are converted into substances such as amyl alcohols and amyl acetates.

Ethylene is consumed in larger amounts for aliphatic petrochemicals than is any other hydrocarbon (about 4,000,000,000 lb/year). About half comes from refinery thermal and catalytic cracking operations conducted primarily to increase the

Table 1. Petrochemicals from methane

Basic derivatives and sources	Produced annually, $\times 10^8$ lb*	Uses*, %
Ammonia	6600	Agricultural chemicals (as ammonia, salts, urea), 76
Petroleum sources		Fibers, plastics, 6
Methane hydrogen, 79%		Industrial explosives, 5
Refinery hydrogen, 4%		Other, 6
Electrolytic, coal, 17%		
Carbon black	1800	Rubber compounding, 96
Natural gas, ~50%		Pigments, metallurgy, 4
Liquid petroleum, ~50%		
Methanol	1700	Formaldehyde (mainly for resins), 36
Petroleum sources		Methanol antifreeze, 6
Methane, 63%		Ethylene glycol, 6
Propane-butane, 7%		Other, 49
Coal, 30%		
Chloromethanes	500	Solvents; cleaners; chlorofluorocarbons for refrigerants, aerosols
Methane chlorination, 50%		
Other sources, 50%		
Acetylene	640	Vinyl chloride, vinyl acetate, 30
Petroleum (mainly methane), ~35%		Chloroprene (neoprene), 26
Calcium carbide, balance		Chloroethylenes, 12
		Acrylonitrile, 8
		Other, 24
Hydrogen cyanide	~150	Acrylonitrile (with ethylene oxide or acetylene), 62
Dept. of Commerce data, proportion from methane probably, >70%		Adiponitrile, 27
		Other, 11

\* Latest figures available.

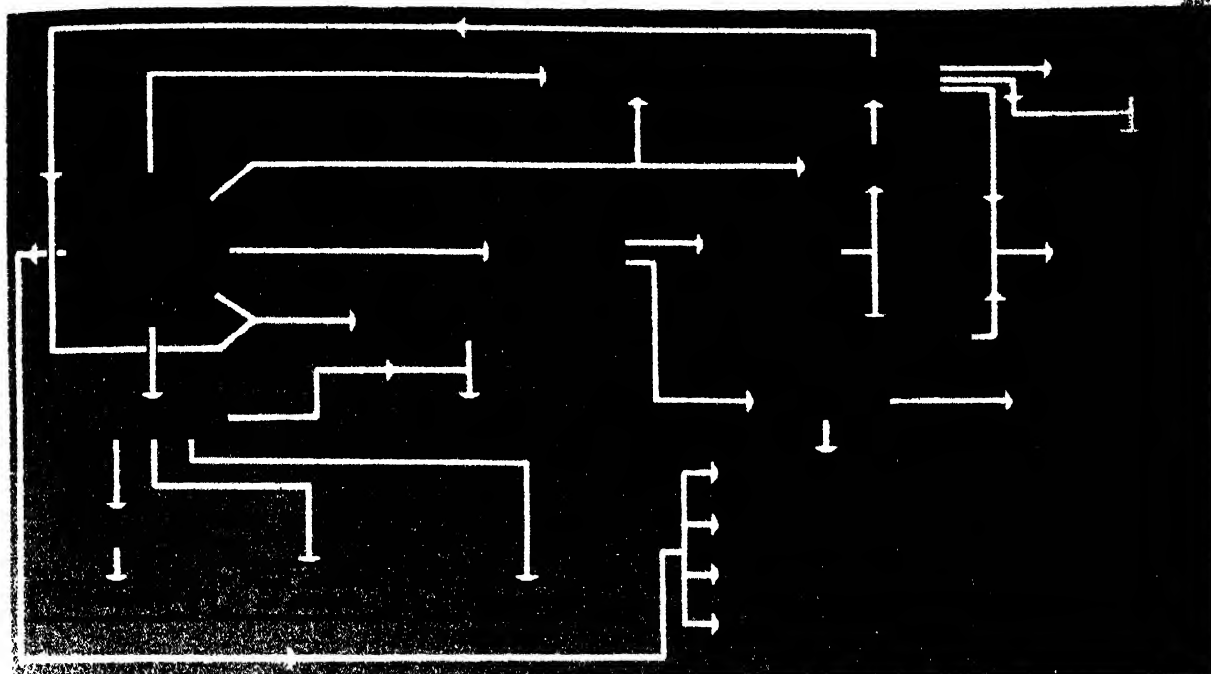


Fig. 3. Petrochemicals from methane.

quantity and quality of gasoline produced, and the rest comes mainly from the cracking of propane and ethane. The principal petrochemicals derived from ethylene are shown in Fig. 4 and Table 2.

Alternative routes from a single petroleum and natural gas source to a certain chemical are illustrated in Fig. 3 (carbon black, ammonia, methyl chloride) and Fig. 4 (ethylene oxide, ethyl alcohol). Alternative petroleum sources are illustrated there also: vinyl chloride and acrylonitrile currently are produced commercially from both acetylene and ethylene. Dovetailing operations are illustrated in Fig. 3 by the synthesis of urea, using carbon dioxide coproduced with the hydrogen required for the ammonia, and in Table 2 by the production and simultaneous utilization of hydrogen chloride in the manufacture of ethyl chloride and of vinyl chloride.

Propylene is produced in large amounts (about 20,000,000,000 lb/year), almost entirely from petroleum cracking primarily for gasoline, and by far the bulk is converted by polymerization or alkylation into high-octane gasoline components. Of the almost 2,000,000,000 lb of propylene consumed annually for chemicals, only 6% comes from cracking done specifically for chemical raw material. The principal petrochemicals derived from propylene are shown in Fig. 5 and Table 3. Of these, isopropyl alcohol is by far the largest in volume, and the only one approaching in scale the several largest ethylene derivatives.

New large-scale derivatives of propylene include polypropylene and hydrogen peroxide. Acrolein, produced on a small scale for several years from petroleum by indirect routes, can be made by direct oxidation of propylene and can be used in part with

petroleum-derived hydrogen peroxide in a second route to synthetic glycerol. The allyl chloride utilized in the original route, also is employed in making epichlorohydrin for epoxy resins (Figs. 5 and 6).

Table 2. Petrochemicals from ethylene

Basic derivatives and sources	Produced annually, $\times 10^6$ lb*	Uses*, %
Ethylene oxide Via direct oxidation, (newer plants), 48 % Via chlorohydrin, 52 %	1240	Ethylene glycol, 66 Di- and triethylene glycols, 9 Ethanolamines, 8 Nonionic detergents, 7 Acrylonitrile, 5 Other, 5
Ethyl alcohol (industrial) From ethylene, 85 % By fermentation, 15 %	1600	Aldehydes, 45 Other chemicals, 25 Solvent, 24 Other, 6
Polyethylene Rapidly growing, capacity about $1500 \times 10^6$ lb	700	Film, sheet, molding, and extrusion
Styrene From ethylene and benzene (or reformate)	1170	Polystyrene, 40 Styrene-butadiene rubber and latex, 48 Misc. resins, 9 Other, 3
Ethyl chloride† Ethylene + HCl, ~88 % Chlorination of ethane, ~12 %	600	Tetraethyllead, ~80 Minor amounts for other ethylations (such as ethyl cellulose)
Ethylene dichloride By direct chlorination, 90 % By-product of ethylene chlorohydrin production, 10 %	800	Vinyl chloride, 70 Scavenger in anti-knock fluid, 20 Other, 10
Ethylene dibromide	160	Scavenger in anti-knock fluid

\* Latest figures available.

† In one process, hydrogen chloride from ethane chlorination is used as reagent for production of ethyl chloride from ethylene.

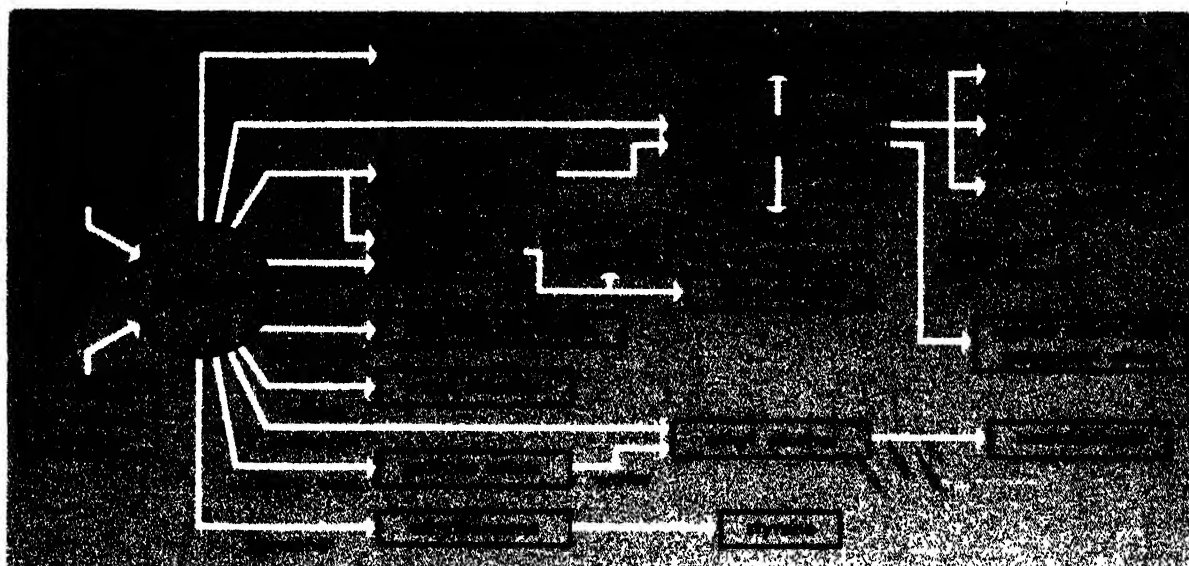


Fig. 4. Petrochemicals from ethylene.

A striking example of the impact of petrochemicals on both natural products and other synthetics is the capture of more than three-quarters of the soap market by synthetic detergents, especially the

**Table 3. Petrochemicals from propylene and the butylenes**

Basic derivatives and sources	Produced annually, $\times 10^6$ lb*	Uses*, %
Isopropyl alcohol From propylene, ~100%	1200	Acetone (by dehydrogenation or oxidation), 46
Propane-butane oxidation, small		Other chemicals, 13
Dodecene (tetramer)	340	Solvent and misc., 41
Nonene (trimer)	90	Detergents dodecylbenzene sulfonate, nonylphenyl polyglycol ethers
Cumene	170	Phenol and acetone
Allyl chloride	100	Glycerol, epichlorohydrin, allyl alcohol
Propylene oxide	80	Mainly propylene glycol, other chemicals
Polypropylene	Small but growing	Expected: film, sheet, molding, and extrusion
Butadiene From butylenes (refinery cracked gas): 60%	1540	Styrene-butadiene rubber, 90
From butane ad hoc: 30%		Acrylonitrile-butadiene rubber, 3
Misc. hydrocarbon cracking: 10%		Styrene-butadiene paint latex, 3
		Adiponitrile and hexamethylenediamine for nylon, 2
		Other, 2
Secondary butyl alcohol	250	Mainly methyl ethyl ketone
Oxo alcohols (via $C_5$ aldehydes): "isooctyl" alcohol	40	Detergents, plasticizers
Butyl rubber (copolymer of isobutylene and isoprene)	240	Inner tubes, mechanical rubber goods
Diisobutylene	110	Mainly detergents, plasticizers
Triisobutylene	Small	
Polyisobutylene	40	
		Adhesives, sealants, electric insulation, lube oil additives
Tertiary butyl alcohol	30	Solvent

\* Latest figures available.

sulfonate from propylene tetramer-benzene alkylate, with a corresponding drop in the production of glycerol as a by-product of soap manufacture. The gap in glycerol supplies has been filled by the introduction of petrochemical glycerol. Also, the assurance of a steady supply of glycerol from readily available propylene has been an important factor in the growth of alkyd resins.

The butylenes (butenes) are in strong demand for the synthesis of high-octane gasoline. Chemical uses account for about 2,100,000,000 lb annually, of which manufacture of butadiene overshadows all others. Source of the butylenes is predominantly cracking conducted primarily for gasoline, with ad hoc cracking of butane and miscellaneous hydrocarbons a minor but growing factor. The principal derivatives of the normal butylenes and of isobutylene are shown in Fig. 5 and Table 3.

The oxo process (reaction of olefins with carbon monoxide and hydrogen) is used in a few plants. It provides aldehydes from olefins of one less carbon number, and from these by hydrogenation, the corresponding primary alcohols. Also in one case at least (using butyraldehyde from propylene), it provides by oxidation the corresponding acid (butyric, for cellulose acetate butyrate). The principal alcohols obtained are branched chain octyl, from the heptene copolymer of propylene and butylene (Fig. 5); branched chain nonyl, from diisobutylene; and branched chain decyl and tridecyl, from propylene trimer and tetramer respectively.

**Cyclic organic compounds.** The coal tar industry has been the traditional supplier of the cyclic compounds, but since 1940, petroleum has become the chief source of many of them. The principal petroleum-derived cyclics are benzene, toluene, the three xylenes, cyclohexane, and their derivatives (Fig. 6 and Table 4), together with two refinery by-product mixtures, naphthenic acids and alkyl phenols. Figure 6 shows only the more important of the cyclic derivatives; thousands of benzene derivatives, for

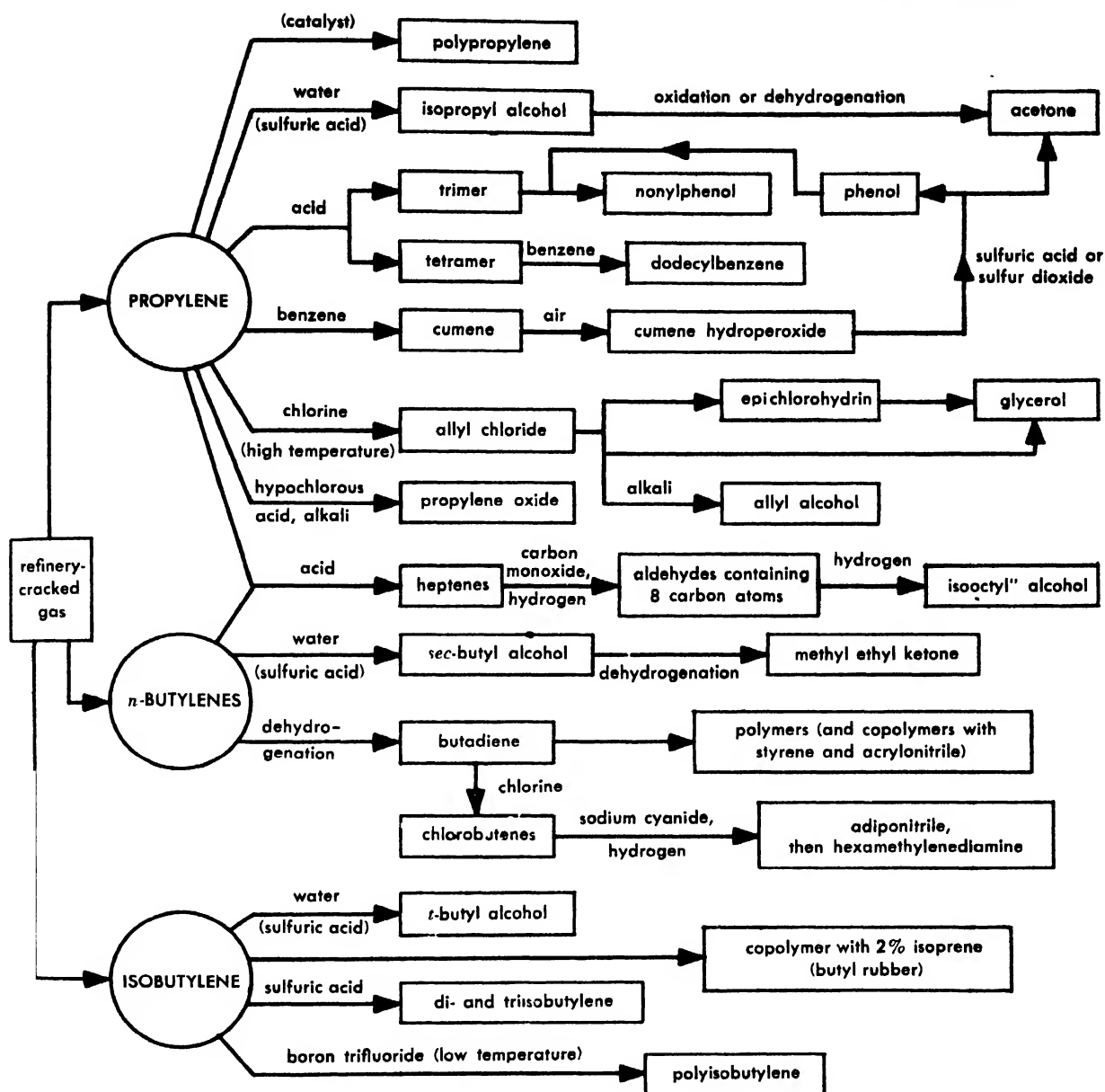


Fig. 5. Petrochemicals from propylene and the butylenes.

example, are known, and hundreds are in commercial production. Those shown are all known to be derived in substantial tonnage from petroleum. High polymers (plastics and resins, fibers, and elastomers) comprise the great bulk of the derivatives, with end products including nylon, polystyrene and styrene rubbers, phenolic resins, epoxy resins, polyurethanes from isocyanates, and phthalate polyesters. The aromatic hydrocarbons also are of much importance in direct uses as high anti-knock gasoline components and solvents.

The successful chemical use of petroleum cyclic hydrocarbons has depended strongly on the development of separation techniques adequate to provide the required purity, since the source materials are complex mixtures containing also a variety of close-boiling paraffins and naphthenes. These techniques include combinations of straight fractional distillation, azeotropic distillation, extrac-

tive distillation, solvent extraction, solid adsorption, crystallization, and chemical conversion.

Naphthenic acids are carboxylic acids of substituted cyclopentanes and cyclohexanes; they are present in crude oils. Commercial naphthenic acids lie in the molecular-weight range 180–350. Owing to their oil solubility, naphthenates of appropriate metals are used as paint driers, fungicides, and lubricant additives. Naphthenic acid production is about 16,000,000 lb annually. Alkyl phenols (cresylic acids) are found in cracked petroleum; the lower boiling of the commercial fractions contain xylenols and smaller amounts of cresols, ethylphenols, trimethylphenols and methylethylphenols. They are richer in ortho and para isomers than coal tar phenols, which are produced at higher cracking temperatures. Their chief uses are in ore flotation frothers, disinfectants, oil and gasoline additives, engine and metal cleaners, and phenolic resins.

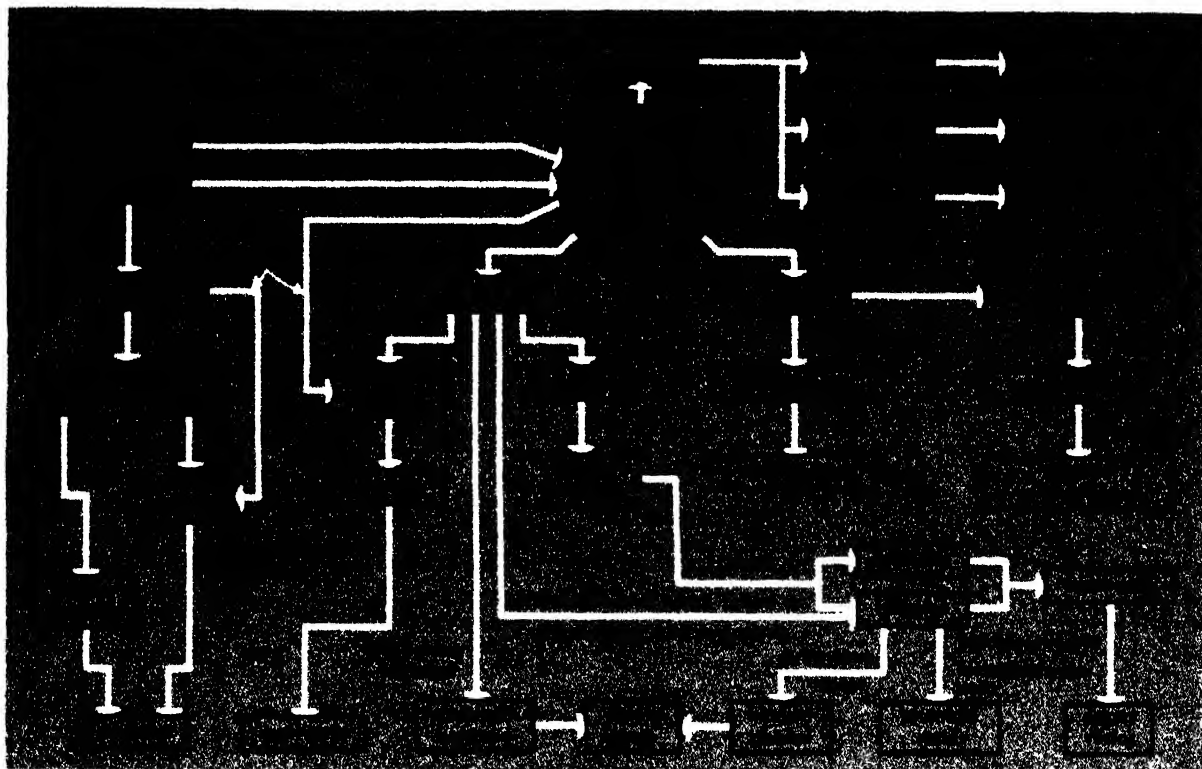


Fig. 6. Cyclic petrochemicals.

Production from petroleum is about 36,000,000 lb annually.

Table 4. Cyclic petrochemicals

Basic products and sources	Produced annually, $\times 10^6$ lb*	Uses*, %
<b>Benzene</b> From petroleum, 35 % From coal, 65 %	2420	Styrene, 41 Phenol via cumene (definitely petrochemical), 4 Phenol via sulfonation or chlorination, 17 Synthetic detergents, 7 Misc. chemicals and other uses, 31
<b>Toluene</b> From petroleum, 79 % From coal, 21 %	1420 (of this, 50 % for non-gasoline uses)	Motor and aviation gasoline, 48 Explosives, 8 Toluene diisocyanate, vinyltoluene, other chemicals, 11 Solvent for coatings, 11 Other, 22
<b>Xylenes</b> From petroleum, 90 % From coal, 10 %	920 (70 % for non-gasoline uses)	Coatings and solvents, 48 Aviation gasoline, 29 Isomer separation (for phthalic acids), 11 Other, 12
<b>Ethylbenzene</b> From benzene, predominant Direct from reformat, small	1260	Styrene, ~100
<b>Cyclohexane</b> From petroleum naphthas, >65 % From benzene by hydrogenation, <35 %	460	Polyamidest (nylons), ~100

\* Latest figures available.

† The hexamethylenediamine required comes mostly from petroleum via adipic acid or butadiene.

**Petrochemical sulfur.** Sulfur is obtained by the oxidation of hydrogen sulfide, which occurs in some natural gases and most refinery gases, particularly in the off-gas from processes to reduce the sulfur content of petroleum liquids (for example, hydrodesulfurization). Petrochemical sulfur production is about 1,440,000,000 lb annually (elemental, from refinery gases, 590,000,000; elemental, from natural gas, 660,000,000; recovered as sulfuric acid from refinery gases, 190,000,000 on a sulfur basis). This is about 8% of all sulfur production in the United States.

The total production of all petrochemicals shown as basic products or basic derivatives in the tables and text is about 24,000,000,000 lb. The 40,000,000,000 lb mentioned earlier and shown in Fig. 2 is a gross amount, in which, owing to the nature of the statistical information, some atoms are counted more than once as they appear in successive important products. The total effort expended in petrochemical manufacture is probably better represented by this gross amount. See PETROLEUM PROCESSING. [H.G.V.; T.W.E.]

**Bibliography:** W. L. Faith, D. B. Keyes, and R. L. Clark, *Industrial Chemicals*, 2d ed., 1957; R. F. Goldstein, *The Petroleum Chemicals Industry*, 2d ed., 1958.

## Petrofabric analysis

A statistical analysis used to determine the arrangement in space (orientation) of rock structures, both on a large scale in rock masses, such as planar and linear structures in sedimentary

rocks, lava flows, igneous intrusions, and stream deposits, and also on a small scale in internal constituents of rocks, such as pebbles, included fragments, and mineral grains, oriented either by shape or by their crystal lattice structure.

Any movement of a natural object is recorded in the position that the object assumes in response to the movement. Preferred orientation, which is a statistical preference of numerous objects for one position in contrast to a random arrangement in space, indicates the direction and type of movement that caused the preferred arrangement. Therefore fabric analysis may lead to a recognition of the kinematic processes by which rocks are formed or deformed.

**Field measurements.** In the field all visible structures of the exposed rock in a given area are measured in as many outcrops as possible. The result of these field measurements is a census of the trend and inclination of various planar and linear structures. These measurements are then plotted on a Schmidt net, which is an equal-area circular projection net of 10-cm radius. The planar structures are plotted as great circles. Linear structures (axes of intersecting planes or axes of folds) are plotted as points. The statistical significance of

these measurements is emphasized by drawing contour lines (isopleths) around the areas of equal density of distribution of points (see illustration). The pattern of preferred orientation will thus become evident as areas of maximum concentration. Patterns formed by individual sets of measurements from several parts of one outcrop, or from one outcrop to another, are compared to determine the uniformity of pattern throughout the area investigated. Fabric diagrams that show uniformity of pattern are combined into one statistical aggregate. The resulting fabric diagram is then reduced in size sufficiently to insert it at the correct spot on the geologic map of the whole area in order to show the uniform structure for that part of the area. A map thus prepared shows the areas that have a similar type of structure in contrast with other areas having a different type.

Superposed structures occur in many regions of strongly deformed rocks. These are represented by diagrams showing different trends of fold axes, superposed in one and the same diagram. Such superposed structures, known as overprints, can then be further studied by a geometrical rotation of the diagram by which the inclinations are flattened out to a horizontal position, a process known as unrolling. The rotation is carried out around the axis formed by the most recent deforming movements. Underlying structures produced by earlier movements can then be recognized.

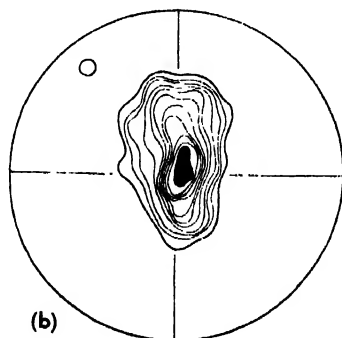
**Study of hand specimens.** In regions of obscure or complex rock structures, representative hand specimens are collected from representative exposures after the geographic coordinates have been marked on the specimens in such a way that each specimen can be set up in the laboratory in exactly the same position that it had in the field. By the use of suitable instruments, structures can then be measured in three dimensions, which in some rocks give a more satisfactory structural pattern than that obtained from field data, which are commonly two-dimensional.

**Study of thin sections.** The study of thin sections under the microscope gives still further valuable information about movements recorded in the rock fabric. Arrangement of the rock-making minerals, by shape and also by crystal lattice structure, is determined by means of a petrographic microscope equipped with a universal stage. Diagrams prepared by this technique give much unsuspected information about the effects of movement in rock fabric. In rocks too fine-grained to study under the microscope, x-ray analysis can give the desired information. [E.B.K.]

**Bibliography:** H. W. Fairbairn, *Structural Petrology of Deformed Rocks*, 2d ed., 1949; E. B. Knopf and E. Ingerson, *Structural Petrology*, Geol. Soc. Am. Mem. 6, 1938.

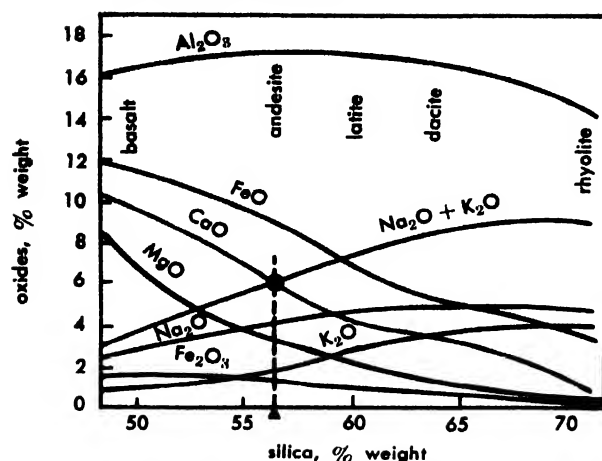
## Petrographic province

A region in which the igneous rocks, formed during a limited period, show a certain community of character (chemical, mineralogical, or petrolog-



Petrofabric diagrams. (a) Random arrangement of field-spar grains in an igneous rock. (b) High degree of preferred orientation of crystallographic *c* axes in quartz in a sheared granite. Out of 158 *c* axes in the quartz, practically all lie within the contoured area of the diagram, the maximum concentration being within the black area in the center. (From B. Sander, *Gefügekunde der Gesteine*, Springer Verlag, 1930)





Igneous rock variation diagram. Representing the calc-alkali series of volcanic rocks, from San Francisco Mountains, Arizona.

ical) which distinguishes them from other rocks in the area. Rocks of a petrographic province are consanguineous in that they are, theoretically at least, derived from a common parental magma. A classic province is represented by the highly potassic rocks around Rome and Naples. Other provinces are characterized by either high or low content in other elements (such as sodium and titanium).

If the chemical analyses of rocks representing a genetically related series are plotted on a variation diagram, many characteristics of the series may be brought out. The accompanying figure shows the percentages of oxides in each rock analysis as plotted against the silica content of that analysis.

All points lie on or close to smooth curves which represent the variation in composition with silica content (or roughly with time because younger rocks tend to be more silicic). Noteworthy are the positive slopes for potassium oxide ( $K_2O$ ) and sodium oxide ( $Na_2O$ ), the arched curve for aluminum oxide ( $Al_2O_3$ ), and the negative slopes for magnesium oxide ( $MgO$ ), calcium oxide ( $CaO$ ), ferrous oxide ( $FeO$ ), and ferric oxide ( $Fe_2O_3$ ). The combined alkali ( $Na_2O + K_2O$ ) curve is seen to cross the  $CaO$  curve at 56.5% silica. This silica value is known as the lime-alkali index for the rock series. On the basis of lime-alkali indices, the numerous rock series are arbitrarily divided into groups as follows: alkalic  $< 51$ , alkali-calcic  $51 < 56$ , calc-alkalic  $56 < 61$ , and calcic  $> 61$ .

Certain rock series appear related to certain types of geological environments. In general the more alkalic types occur in regions subjected to tension and vertical movement (faulting and subsidence), whereas more calcic series are found in compressional regions (fold-mountain belts).

**The olivine basalt-trachyte association.** Rocks belonging to this association are extremely widespread. The association is well represented in the central Pacific islands (Hawaii, Tahiti, Samoa); the islands along the mid-Atlantic ridge (Ascension,

Saint Helena); and islands of the Indian Ocean (Kerguelen). Both mineralogically and petrographically the association is relatively simple. Primary olivine basalt magma has given rise to the following sequence largely through crystallization and sinking of heavy minerals.

Olivine basalt  $\rightarrow$  basalt  $\rightarrow$  andesite  $\rightarrow$  trachyte

The last two types are not abundant. Locally the series may be carried beyond trachyte to quartz trachyte (Samoa) or soda rhyolite (Ascension). Elsewhere it may pass through tephrite to phonolite. The extrusion of phonolite may be due to a strongly undersaturated (silica-deficient) parent magma. The quartz-bearing end products may be due in some areas to a saturated parent magma and in others (Ascension) to slight assimilation of older granitic rocks.

The characteristic development of oceanitic, ankaramite, and limburgite may be explained by gravitational accumulation of olivine and pyroxene at lower levels in the volcanic reservoir. Density stratification of this type would permit nearly simultaneous eruption of highly contrasting lavas from neighboring vents.

The olivine basalt trachyte association is well represented on the continents (Otago, New Zealand; Oslo, Norway; East African rift zone; Midland Valley of Scotland). Here, however, the later members of the series (trachyte, soda rhyolite, and phonolite) are relatively more abundant; and leucite (rare in oceanic areas) may be locally important.

In general, the association occurs in regions of faulting and marked vertical movement. See MAGMA.

**Flood basalts.** These basalts, also known as plateau basalts, form thick accumulations of nearly horizontal flows over vast areas (hundreds of thousands of square miles). Examples include the Columbia-Snake River basalts of Oregon and Washington, the Deccan plateau lavas of western India, and the Keweenaw lavas of Lake Superior. These flows were probably erupted through fissures from great deep-seated supplies of basaltic magma.

Rocks of this association are overwhelmingly basalts, but rare amounts of rhyolite, trachyte, and andesite occur. Some geologists distinguish two types of flood basalts, olivine basalt and tholeiitic basalt. The former is slightly lower in silica but higher in soda, potash, and magnesia than is the latter. Both types may occur in the same area. They may be derived from different earth shells (at different levels), or the tholeiitic type may form from olivine basalt by crystal fractionation. Other geologists consider the two types merely variations of a single primary basalt magma. See BASALT.

Intrusive masses (sills and dikes) of basaltic material (diabase) form swarms over tens of thousands of square miles in many parts of the world. Particularly notable are the nearly flat sheets of diabase in the sediments of the Karroo system in

South Africa and the flat Palisade sill of New Jersey. The thick, tabular bodies of magma differentiated somewhat as they cooled; and olivine accumulated in a layer near the sill floors. Pyroxene is less abundant and more iron-rich toward the top, and plagioclase increases in abundance and soda content in the same direction. These relations demonstrate the control of crystal fractionation and settling in the process of differentiation. *See* DIABASE.

**Basalt-andesite-rhyolite association.** The andesite-rhyolite kindred appears on continents and is typical of but not restricted to regions of orogeny (fold-mountain belts). Most striking is the circum-Pacific belt of volcanoes (many still active) extensively developed along the western margin of North and South America.

The rocks consist chiefly of andesite and rhyolite with smaller amounts of basalt, dacite, latite, and quartz latite. Both lavas and tuffs are represented. The sequence of eruption of rock types is extremely complicated and varies with location as well as with time. Basaltic magma may have been the parental fluid in most areas. Differentiation of this basic melt has operated by fractional crystallization, but superimposed upon its effects are those of assimilation of crustal rocks and mixing of partly crystallized magmas (consanguineous) at various stages of differentiation. Possibly much rhyolitic and andesitic magma was generated by local melting of the crustal rocks.

**Basic-ultrabasic association.** This association includes rocks of basic or ultrabasic composition and is found in large sheetlike, saucer-shaped, or conical bodies commonly injected at shallow depths. Examples include the Duluth lopolith of Minnesota, the Stillwater complex of Montana, the Skaergaard body of east Greenland, and the Bushveld complex of South Africa.

The rock types are distributed in flat or somewhat basined layers to form sequences thousands of feet thick. In general the rocks become heavier toward the bottom, but in detail, thin layers of contrasting types may alternate to give a strongly banded appearance. Peridotite and pyroxenite are most abundant near the base. Upward these give way to norite and gabbro with some anorthosite. Uppermost rocks may be dioritic to granitic. Olivine and pyroxene become richer in iron, and plagioclase becomes more sodic from bottom to top in these bodies.

Such distributions suggest that the originally injected basaltic or basic magma solidified in general from the floor upward. Crystals may have initially formed in upper regions and may have settled to build up the floor. Convection currents in the melt may have helped to redistribute and sort the crystals of different size and density. The abundant granitic material commonly found near the roof may represent, in varying amounts, a late differentiation of the original basaltic magma, a product of assimilation of siliceous rocks (sediments and felsites) by the magma, or a crystallized secondary melt generated by the hot basic intrusion.

**Granodiorite-granite association.** This extensively developed, coarse-grained plutonic assemblage is restricted to continents and may be subdivided into two categories. The first includes the smaller masses (stocks, ring-dikes, etc.) commonly widely scattered and formed at relatively shallow depths. These masses usually transgress the structure of surrounding rocks, but some are highly concordant and may have spread or domed the adjacent and overlying rocks. They are usually surrounded by a metamorphic halo or zone of recrystallization which is generally most conspicuous where the rocks have not been previously metamorphosed.

Granite and granodiorite predominate, with minor amounts of diorite, gabbro, and syenite present. Most bodies have probably formed by crystallization of granite or granodiorite magma. Where minor basaltic melts have been involved, crystal fractionation and assimilation of adjacent rock material may have been operative. *See* AUREOLE, CONTACT.

The second category includes immense, deep-seated bodies (batholiths) surrounded by metamorphic rocks and restricted to orogenic zones. The Sierra Nevada batholith of California and the Coast Range batholith of British Columbia are typical examples. The predominant rock type is granite in those masses of Precambrian age and granodiorite in the younger bodies. Quartz diorite is somewhat less abundant whereas diorite, gabbro, and syenite occur only locally. Batholiths are generally elongate parallel to the orogenic belt, but locally they are highly crosscutting. The granitic rocks may be massive or foliated (showing parallel streaking or layering of minerals) much like the adjacent metamorphic rocks into which they may appear to grade. Many are characteristically associated with pegmatites and migmatites.

Some of these large bodies may form from granitic magmas derived by downbuckling and melting of the earth's sialic (granitic) layer. Others may represent more or less reconstituted (metamorphosed) sediments in geosynclines. Still other bodies may be products of granitization. *See* GRANITIZATION; METAMORPHIC ROCKS; METAMORPHISM; MIGMATITE; PEGMATITE.

**Leucite basalt-potash trachybasalt.** This association includes a wide variety of silica-poor, potash-rich rocks of volcanic and near surface origin. The association is confined to the continents. Basic (low in Si, high in Ca, Fe, and Mg) and ultrabasic lavas with leucite are dominant in many areas. Occurrences are restricted but widespread (Rome-Naples, Italy; Uganda, east Africa; West Kimberley, Western Australia; and Leucite Hills, Wyoming). Rock types include leucite basalt, leucite basanite, potash trachybasalt, and melilitite basalt. The association appears in regions of faulting and marked vertical movement. *See* LEUCITE ROCK.

**Spillite-keratophyre association.** This association includes volcanic flows and tuffs with minor intrusives, intimately associated with sediments in

geosynclinal regions. The rocks are soda-rich and potash-poor; they are chiefly basaltic (spilites) with some soda trachyte (keratophyre). Many appear altered (metamorphosed), and some are associated with rocks of the basalt-andesite-rhyolite suite. See SPILITE.

**Ultrabasic rock.** The predominantly peridotite and serpentine rock in abundant intrusive bodies is closely associated with the spilitic suite. Together the two rock associations constitute the so-called ophiolites, generally considered to represent the earliest magmatic phenomenon in orogenic regions. See PERIDOTITE.

**Nepheline syenite association.** Nepheline syenite and associated alkali-rich rocks are widespread but rare. They are continental rocks and commonly appear in areas of subsidence and faulting. See NEPHELINE SYENITE.

**Anorthosite.** This rock forms gigantic masses in Precambrian terranes and appears associated with hypersthene granite and norite. It is composed of andesine or labradorite and, therefore, differs from the calcium-rich anorthosite associated with gabbro in large stratiform sheets. See GABBRO.

[C.A.C.A.]

**Bibliography:** T. F. W. Barth, *Theoretical Petrology*, 1952; F. J. Turner and J. Verhoogen, *Igneous and Metamorphic Petrology*, 1951.

## Petrography

A branch of petrology, the study of rocks, that emphasizes the description and systematic classification of rocks, especially by the study of thin sections under the petrographic microscope, by analysis of disaggregated samples, or by analysis of individual mineral components. See PETROLOGY.

The megascopic classification of rocks, based on characteristics observed in field exposures or hand specimens, suffices for some purposes; but refined description and classification require determination of the kinds, sizes, shapes, and space interrelationships in the aggregates of original mineral components and the changes and alterations that have affected rocks or their separate components subsequent to their initial formation.

The petrographic investigation of a rock usually entails both an analysis of the rock after disaggregation into mineral or particle size fractions and examination of the rock in thin sections under the petrographic microscope.

**Study of disaggregated samples.** Techniques for study of disaggregated samples differ depending upon whether the rock is composed of an accumulation of particles and fragments as in sandstone (clastic rock) or consists of interlocking mineral grains as in granite (crystalline rock).

Poorly to moderately indurated, or hardened, clastic rocks are disaggregated by crushing or pulverization, by dissolving cementing materials, or by using the disrupting effect of salts precipitating from saturated solutions with which the rock is impregnated. After disaggregation the size distribution of the clastic particles is ascertained by screen

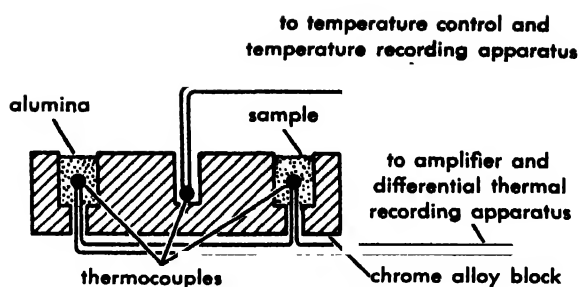


Fig. 1. Idealized diagram showing principal components of sample holder used for differential thermal analysis. (From E. E. Wahlstrom, *Petrographic Mineralogy*, Wiley, 1955)

analysis or by liquid or air elutriation; and the shapes, roundness, and surface characteristics of the particles are noted under a stereoscopic binocular microscope.

Hard clastic rocks and crystalline rocks, including igneous, metamorphic, and recrystallized or firmly cemented sedimentary rocks, are disaggregated by crushing into fragment sizes that will permit separation into mineral fractions. A variety of methods for separation of disaggregated rocks into mineral fractions are available.

**Separation of mineral fractions.** Commonly used methods include hand sorting under the microscope, separation in heavy liquids, magnetic separation, and electrostatic separation.

**Heavy-liquid separation.** Heavy liquids consisting of suspensions of ground metals in liquids are available with densities as high as 7.5 but are infrequently used because of the opacity. Widely used transparent liquids are bromoform, which has a density near 2.9, and acetylene tetrabromide, with a density of 2.96. These are employed to separate heavy minerals from light minerals. Powdered rocks, sometimes sized by screening and washed to remove the very fine fractions, are suspended in a heavy liquid, in a funnel or evaporating dish, and vigorously agitated to cause the light minerals to float and the heavy minerals to sink. After the heavy minerals have been collected and washed, successive crops of the light minerals may be obtained by careful progressive dilution of the heavy liquid with its appropriate solvent.

**Magnetic separation.** Electromagnetic separators provide mineral fractions differing in magnetic susceptibilities. Satisfactory electromagnetic separators are constructed to allow variations in field intensity at the point or area of separation, variation of the tilt of the poles of the magnet, and variation of the rate of feed. Best results are obtained with clean, free-flowing aggregates of uniformly sized grains. Electromagnetic separation is not satisfactory when minerals contain nonuniformly distributed impurities or alteration products.

**Electrostatic separation.** Minerals of different electrical conductivities can be segregated by electrostatic separation. Mineral grains of approximately equal size placed on a grounded metal plate

are differentially attracted to a charged surface, and separation of the different mineral fractions is accomplished by varying the intensity of the charge or the spacing between the charged surface and the plate supporting the mineral grains. As in magnetic separation, irregularly included impurities and alteration products prevent clean separations.

**Study of mineral fractions.** After separation, each mineral component is identified by one of the several techniques of determinative mineralogy (see MINERALOGY). Preliminary identification by physical properties and by simple wet and dry tests using the procedures of systematic blowpipe analysis may precede more elaborate tests. Spectrographic analysis, microchemical tests, or partial or complete chemical analyses are made to identify gross or trace elements. Particularly useful in identification are x-ray diffraction patterns of powdered minerals as recorded on film strips or as traced on charts on recording diffractometers. See CHEMICAL MICROSCOPY; SPECTROCHEMICAL ANALYSIS; X-RAY FLUORESCENCE ANALYSIS.

Minerals such as the feldspars, feldspathoids, carbonates, and clay minerals may be treated chemically so that they assume distinctive colors resulting from absorption of dyes or from precipitation of colored chemical compounds. Minerals that undergo rapid chemical changes or crystallographic inversions during heating are studied successfully with differential thermal apparatus (Fig. 1) or by determination of change of weight during heating in the case where decomposition with loss of a volatile substance takes place. Differential thermal analysis measures the magnitudes and temperatures of endothermal and exothermal reactions in minerals as they are heated in a furnace from room temperature to temperatures of 1000°C or more. See EQUILIBRIUM, PHASE; THERMOCHEMISTRY; TRANSITION POINT.

Rocks consisting in whole or part of clay minerals or clay-sized particles pose special problems. Mineral fractions or unseparated aggregates of clay minerals are studied by stain tests, x-ray diffraction patterns, differential thermal analysis,



Fig. 2. Four-axis universal stage. (E. Leitz, Inc.)

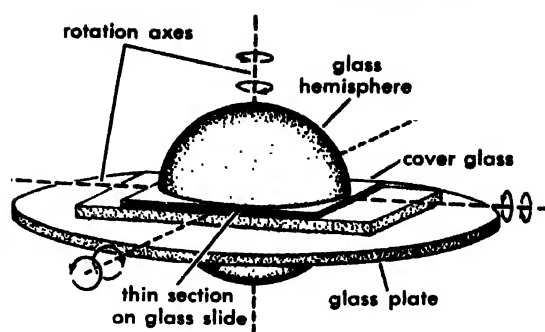


Fig. 3. Thin section mounted between glass hemispheres on universal stage. (Adapted from E. E. Wahlstrom, *Petrographic Mineralogy*, Wiley, 1955)

base exchange properties, infrared absorption, and the electron microscope. See MICROSCOPE, ELECTRON.

**Microscopic petrography.** For rocks in which the mineral components are large enough to be identified under the microscope, examination under a petrographic microscope of thin sections ground to a thickness of 0.03 mm is standard procedure. The petrographic microscope differs from an ordinary microscope in that it has a rotating stage and polarizing filters or prisms which produce optical phenomena not observable in ordinary light. Minerals are identified by their optical and other physical properties by using tables and charts which correlate physical properties with chemical composition.

Minerals previously separated from disaggregated rocks are further analyzed under the microscope by placing fragments in liquid immersion mediums of known refractive indices and determining the refractive index or indices of the mineral by a matching process. A useful aid in optical studies is the universal stage (Fig. 2), a multiaxis device which is attached to the stage of the microscope and permits rotation of thin sections or mineral fragments mounted in immersion mediums into any desired position for the measurement of optical properties (Fig. 3). Thin sections from which the cover glass has been removed may be etched and stained to aid in the identification of some components by the same reagents used for staining crushed fragments.

A particular advantage of the thin-section technique is that it allows measurement of the dimensions of the mineral components and reveals the nature of the contacts and the manner of intergrowth of the mineral components. Subtle differences in composition within various portions of single crystals, twinning, strain, alterations, and other features that might not be noted in nonoptical techniques become very apparent under the microscope.

**Mode.** The mode of a rock expresses the mineral composition in weight or volume percentages. The mode of crushed or otherwise disaggregated rocks is obtained by weighing the separated mineral

fractions. The volume mode of rocks obtained from thin sections may be converted to a weight mode by calculations using assumed or measured specific gravities for each mineral component. Weight modes are used to calculate bulk chemical compositions of rocks after the composition of each mineral has been determined by optical or other means. The volume mode is determined by micrometric analysis in thin sections. With a mechanical stage several linear traverses are made across the thin section and the mode is calculated on the assumption that the volume of each mineral is proportional to the total of the linear intercepts for the mineral measured by the stage. A more rapid method employs a point-counter, a mechanical stage which moves the thin section to a succession of equally spaced points in a linear traverse. The mineral at the intersection of the cross hairs of the microscope is noted at each point, and the volume of each mineral in the rock is assumed to be proportional to the number of points counted for each mineral.

**Space arrangement.** The space arrangement of the mineral components is analyzed quantitatively by means of petrofabric techniques. The angular relationships to the plane of the thin section, and ultimately to the site of collection of the sample in the field, of crystallographic directions, as determined by optical measurements or by observation of crystal shapes, are measured with a universal stage. Measured angles are used to plot points on an equal-area projection, a special type of projection derived from the spherical projection. The density of points on the projection is indicated by contours, so that the finished plot is a statistical indication of the preferred orientation in one respect or another of one or several mineral components. *See* PETROFABRIC ANALYSIS.

A specialized branch of microscopic petrography deals with opaque minerals, especially ore minerals, as studied under a reflecting microscope. The highly polished surface of a mineral or mineral aggregate is examined by reflected light which is directed onto the surface of the mineral by a vertical illuminator, a device inserted into the microscope between the object lens and the barrel of the microscope. Polarizing plates or prisms permit observations in plane-polarized light. Minerals are identified by their isotropism or anisotropism, color, reflectivity, hardness, and response to a standard set of etch reagents. Small portions of minerals scratched from the polished surface are subjected to systematic microchemical analysis, x-ray analysis, or spectrographic analysis. *See* MICROSCOPE, REFLECTING.

Rocks containing both opaque, and nonopaque minerals may be studied by making polished thin sections. Part of a section is left uncovered so that the opaque components can be studied in reflected light while the other components are examined by transmitted light by ordinary methods. [E.E.W.]

**Bibliography:** E. W. Heinrich, *Microscopic Petrography*, 1956; L. W. LeRoy, *Subsurface Geo-*

*logic Methods*, 2d ed., 1950; E. E. Wahlstrom, *Petrographic Mineralogy*, 1955; H. Williams, F. J. Turner, and C. M. Gilbert, *Petrography*, 1954.

## **Petrolatum**

A smooth, semisolid blend of mineral oil with waxes crystallized from the residual type of petroleum lubricating oil. The wax molecules contain from 30 to 70 carbon atoms and are straight chains with a few branches or naphthene rings. They are micro-needles and hold a large amount of oil in a gel. Petrolatums are useful because they cling, lubricate, and resist both moisture and oxidation. They serve as lubricants in baking and candymaking; as carriers in polishes, cosmetics, and ointments; as rust preventives; as waterproofing agents for paper; and in other uses calling for an inert grease-like material (*see* PETROLEUM PRODUCTS). [J.K.R.]

## **Petroleum**

A naturally occurring, oily, flammable liquid composed principally of hydrocarbons, and occasionally found in springs or pools but usually obtained from beneath the earth's surface by drilling wells. Formerly called rock oil, unrefined petroleum is now usually termed crude oil.

Petroleum is separated by distillation into fractions designated as (1) straight-run gasoline, boiling up to about 200°C.; (2) middle distillate, boiling at about 185–345°C. from which are obtained kerosene, heating oils, and diesel, jet, rocket, and gas turbine fuels; (3) wide-cut gas oil, which boils at about 345–540°C. and from which are obtained waxes, lubricating oils, and feed stock for catalytic cracking to gasoline; and (4) residual oil, which may be asphaltic.

The physical properties and chemical composition of petroleum vary markedly depending on its source. As it comes from the earth, it ranges from an occasional nearly colorless liquid consisting chiefly of gasoline to a heavy black tarry material high in asphalt content. Although most crudes are black, many are amber, red, or brown by transmitted light and show a greenish fluorescence by reflected light. Their specific gravity is usually in the range between about 0.82 and 0.95.

Hydrocarbons constitute 50–98% of petroleum, and the remainder is comprised chiefly of organic compounds containing oxygen, nitrogen, or sulfur, and trace amounts of organometallic compounds. Pennsylvania crude oils contain 97–98% hydrocarbons; some California oils contain only 50%.

**Hydrocarbon types.** The hydrocarbon types found in petroleum are paraffins (alkanes), cycloparaffins (naphthenes or cycloalkanes), and aromatics. Olefins (alkenes) and other unsaturated hydrocarbons are usually absent.

**Paraffins.** The paraffins range from methane (found together with ethane, propane, and the butanes in the natural gas which accompanies petroleum) to *n*-hexacontane (C<sub>60</sub>H<sub>122</sub>, a microcrystalline wax) and compounds of even higher molecular weight. Both straight-chain and

branched-chain paraffins are present. The former usually predominate, particularly in the higher-boiling fractions. Commercial paraffin wax ordinarily consists chiefly of straight-chain paraffins of from about 22–30 carbon atoms isolated from the wide-cut, gas-oil fraction. See ALKANE.

**Cycloparaffins.** The cycloparaffins are chiefly those having five or six carbon atoms in the ring. These include not only the monocyclic compounds (cyclopentane, cyclohexane, alkylcyclopentanes, and alkylcyclohexanes) but also polycyclic hydrocarbons such as the bicycloparaffins (*trans*-decahydronaphthalene and *cis*-bicyclo[3:3:0]octane) as well as tri- and higher cycloparaffins. See ALICYCLIC HYDROCARBON.

**Aromatics.** Aromatic hydrocarbons are usually present in smaller amounts than the paraffins and cycloparaffins. The aromatic compounds boiling in the gasoline range are chiefly alkylbenzenes (such as toluene, the xylenes, *p*-cymene). Higher-boiling fractions contain polynuclear aromatics of both fused-ring (alkylnaphthalenes) and linked-ring (biphenyl) types. The fused-ring polycyclics usually predominate. Mono- and polynuclear aromatic rings fused to one or more cycloparaffin rings (as in indan and 1,2,3,4-tetrahydronaphthalene) are also present. See AROMATIC HYDROCARBON.

**Petroleum fractions.** The number of carbon atoms in hydrocarbons of a given boiling range depends on the hydrocarbon type. In general, gasoline will include hydrocarbons having 4–12 carbon atoms; kerosine, 10–14; middle distillate, 12–20; and wide-cut gas oil, 20–36.

A study of the gasoline fractions from representative petroleum from seven different areas in the United States has permitted some interesting conclusions. Five main classes of compounds are present in the gasoline fraction: straight-chain paraffins, branched-chain paraffins, alkylcyclopentanes, alkylcyclohexanes, and alkylbenzenes. Although the relative amounts of the classes vary from petroleum to petroleum, the relative amounts of the individual compounds within a given class are of the same magnitude for the different petroleum. Hence, the gasoline fraction of different crudes is characterized by specifying the relative amounts of the five main classes.

Petroleum may be classified in accordance with their composition. Thus, Pennsylvania and Michigan crude oils are largely paraffinic and contain little or no asphalt. Some Texas and California oils are rich in naphthenes, whereas others are unusually high in aromatics; most contain much asphalt.

Asphalt is a dark brown to black solid or semi-solid consisting of carbon, hydrogen, oxygen, sulfur, and sometimes nitrogen. It is made up of three components: (1) asphaltene, a hard, friable, infusible powder; (2) resin, a semisolid to solid ductile and adhesive material; and (3) oil, which is structurally similar to the lubricating oil fraction from which it is derived. The asphalts are almost completely soluble in carbon disulfide, carbon tetrachloride, and pyridine but are only partly solu-

ble in low-boiling paraffins which dissolve the oils and resins and precipitate the asphaltenes. See ASPHALT AND ASPHALTITE.

**Components other than hydrocarbons.** The total oxygen content of crude oils is generally low but may be as high as 2%. The oxygen-containing compounds consist principally of phenols and carboxylic acids. The phenols comprise cresols and higher-boiling alkylphenols. The acids include straight-chain and branched-chain acids such as hexanoic acid and 3-methylpentanoic acid, and cyclopentane and cyclohexane derivatives such as cyclopentaneacetic acid and cyclohexanecarboxylic acid. There is also some indication of the presence of acids containing aromatic rings (mono- and dinuclear). Hence, the name naphthenic acids which has been applied to the carboxylic acids derived from petroleum is a misnomer; petroleum acids is a preferable term.

The nitrogen content of crude oils ranges from less than 0.05 to about 0.8%. Up to about one-half is in the form of basic pyridine and quinoline compounds, the latter predominating. The nonbasic nitrogen compounds or complexes include pyrroles, indoles, and carbazoles.

The sulfur content varies over a wide range, from traces to more than 5%. Pennsylvania and midcontinent crudes usually contain less than 0.25% by weight of sulfur, whereas some California and Texas stocks contain over 2%. Part of the sulfur may be in the form of elemental sulfur and hydrogen sulfide. Most is present as mercaptans (thiols), aliphatic sulfides, and cyclic sulfides. The mercaptans and sulfides exist as both straight-chain compounds such as *n*-propyl mercaptan and methyl ethyl sulfide, and branched-chain compounds such as *tert*-butyl mercaptan and methyl isopropyl sulfide. The cyclic sulfides consist of five- and six-membered ring compounds such as thiacyclopentanes and thiacyclohexanes.

A number of metals have been identified in the ash (about 0.01 to about 0.05% by weight) obtained by burning crude petroleum. These include sodium, magnesium, calcium, strontium, copper, silver, gold, aluminum, tin, lead, vanadium, chromium, manganese, iron, cobalt, nickel, platinum, and uranium. Boron, silicon, and phosphorus have also been detected. It is quite probable that the sodium and strontium are present chiefly in the form of aqueous solutions of salts that are finely dispersed in the oil. Most of the other metals are present as oil-soluble salts or organometallic compounds. For example, nickel and vanadium which are the most abundant of these, occurring in 5–40 parts per million in many crude oils of the United States, are probably present as porphyrin complexes. See PETROCHEMICAL; PETROLEUM (ORIGIN); PETROLEUM ENGINEERING; PETROLEUM GEOLOGY; PETROLEUM PROCESSING; PETROLEUM PRODUCTS; PROSPECTING, PETROLEUM. [L.S.]

**Bibliography:** B. T. Brooks et al. (eds.), *The Chemistry of Petroleum Hydrocarbons*, vol. 1, 1954; A. E. Dunstan et al. (eds.), *The Science of*



*Petroleum*, vol. 2, 1938; vol. 5, pt. 1, 1950; H. L. Lochte and E. R. Littmann, *The Petroleum Acids and Bases*, 1955; F. D. Rossini, B. J. Mair, and A. J. Streiff, *Hydrocarbons from Petroleum*, 1953.

## **Petroleum (origin)**

Petroleum, a complex mixture of hydrocarbons, contains small amounts of oxygen, nitrogen, and sulfur compounds, and traces of metal salts. Accumulation of petroleum is believed to involve three steps: generation of oil, primary migration (the movement of oil from source to reservoir rock), and secondary migration (the redistribution of oil within the reservoir rock to form a pool). On the basis of the best available data, these steps can be described as follows.

**Generation of oil.** Oil is generated in sedimentary basins. These basins are shallow continental depressions, hundreds of square miles in area, that have intermittently been covered with sea water and are now filled with sediments. The sediments are of three types: (1) rock particles varying from sands to clay muds, which were eroded from hills and mountains and carried to the basins by streams; (2) biochemical and chemical precipitates such as limestone, gypsum, anhydrite, and chert; and (3) organic matter from the plants and animals that lived in the sea or were carried in by rivers. Some of these types of sediments are being laid down today in basins such as the Persian Gulf and the Caspian Sea.

The third type of sediment, the organic matter, is considered the source of petroleum. Evidence for this is the fact that petroleum contains traces of several substances that could have come only from living things. Examples of these are porphyrins related to hemin and chlorophyll, which are components of modern organisms; optically active compounds (compounds that will rotate the plane of a ray of polarized light); and structures related to cholesterol.

It is believed that oil is generated from organic matter in one or both of two ways. It may come directly from the hydrocarbons that marine organisms form as part of their living cells, or it may come from the conversion of dead organic matter to hydrocarbons.

Evidence of the first process is that some of the complex hydrocarbons in petroleum are comparable to those found in modern plankton, kelp, other algae, coral, and higher organisms such as oysters and bluefish. When such organisms die in the waters of a sedimentary basin, their remains drop into the material accumulating in the basin. Their hydrocarbon content is low (0.005–0.1% by weight), and most of this is destroyed by bacterial oxidation. However, the total amount of hydrocarbon produced in this manner is so great (probably more than 1,000,000 barrels a year) that less than 1% of it would have to be preserved and accumulated to account for a possible  $1.5 \times 10^{12}$  bbl of recoverable oil existing today in the land and continental shelf sediments.

The second process by which oil may be formed involves the synthesis of hydrocarbons from the decay and alteration of buried organic matter. It is believed that bacteria convert the organic matter into more petroleum-like substances by removing oxygen, sulfur, and nitrogen in the form of water, hydrogen sulfide, and ammonia, respectively. The earth's crust contains such a vast amount of organic matter that if only traces of it were converted in this manner it could form all the world's oil.

The synthesis process probably occurs over relatively short periods of time; there are many oil accumulations that appear to have formed from rocks less than 3,000,000 years old. In fact, since hydrocarbon content does not increase with depth in the rocks of sedimentary basins, synthesis may be essentially completed in the first few hundred feet of burial. The temperatures under which the process takes place are probably low, because the temperatures of petroleum source and reservoir rocks rarely exceed 100°C.

A particular environment is required to preserve the biogenetically produced hydrocarbons and to make conversion of organic matter possible. The source beds of petroleum are fine-grained clay or carbonate muds deposited in basins under reducing conditions. Coarse sediments such as clean sandstones, reefs, and oolites are usually not source rocks of petroleum because they are deposited in shallow-water areas, where water movement winnows out the organic matter and the oxygen-containing environment tends to destroy hydrocarbons.

Present-day basins provide conditions that may be representative of those under which oil can be formed, and that can be studied. For example, recent sediments in the Gulf of Mexico contain up to 0.05% of young (less than 15,000-year-old) hydrocarbons that are of the same general types found in petroleum, such as paraffins and naphthenes. However, certain differences in the structure and distribution of these types suggest that some of them are an intermediate stage between hydrocarbons from organisms and petroleum.

**Primary migration.** Primary migration of petroleum from source to reservoir, the second step in accumulation, is caused by the movement of water, which carries oil in concentrations of less than 100 parts per million out of compacting sediments.

When the source muds are deposited, they contain 70–80% water. The remainder is solids, mostly clay minerals or carbonate particles; as they build up to great thicknesses in a sedimentary basin, water is squeezed out by the weight of the overlying sediments. After compaction the source muds contain only about 25% water at a depth of 2000 ft, and 10% at 6000 ft.

The water and oil move in the direction of least resistance (lowest hydrostatic pressure) at a rate of about 1–3 in. per year. Early in compaction, the direction of fluid movement is straight up. As compaction progresses, there is lateral as well as vertical movement of the fluids. The lateral movement results primarily from the tendency of the flat



clay mineral particles to lie horizontally as they are compressed. This reduces the vertical permeability of the compacting muds. In addition, the long, continuous sands on the edges of basins orient fluid movement laterally as burial progresses.

The mechanism by which the water carries the oil is uncertain. It may travel in any of three ways—in solution, as a colloidal dispersion, or as droplets of oil. Transportation in solution is improbable because the solubility of most petroleum hydrocarbons in water is too low to account for known oil accumulations. The droplet hypothesis is improbable because the oil would have difficulty in moving through the fine pore openings in the shale. The oil probably travels as a colloidal dispersion of oil in water, stabilized by natural dispersing agents. When this dispersion reaches the reservoir its salinity is lowered, and this change in environment may cause coagulation of the colloidal hydrocarbons to form discrete oil particles.

If the porous rock is completely enclosed with compacting muds, the oil particles will be held in by the capillary pressures in the muds and the water will pass through. If there is a porous avenue of escape that bypasses the muds, hydraulic currents created by the moving water will sweep the oil droplets out of the reservoir. As compaction progresses and most of the water is expelled from the muds, they will develop into tight shales or dense carbonates. These form the seal for the accumulated oil particles.

**Secondary migration.** In secondary migration, the last stage in the origin of an oil accumulation, the oil droplets are moved about within the reservoir to form the pool. Secondary migration includes in some instances a second step, during which crustal movements of the earth shift the position of the pool within the reservoir rock.

The position of the accumulating pool is affected by several, sometimes conflicting, factors. Buoyancy causes oil to seek the highest permeable part of the reservoir; capillary forces direct the oil into the coarsest grained portion first, and into successively finer-grained portions as the reservoir is filled. Any permeability barriers in the reservoir channel the oil into a somewhat random distribution. Oil accumulations in carbonate rocks are often erratic because part of the original void spaces have been plugged by minerals introduced from water solutions after the rock is formed. In large sand bodies, barriers formed by thin layers of dense shale may hold the oil at various levels. When crustal movements of the earth occur, oil pools are sometimes shifted away from the place in which they originally accumulated. Faults sometimes cut through reservoirs, destroying part of the pools or shifting them to different depths. Uplift and erosion bring the pools near the surface, where the lighter hydrocarbons evaporate. Fracturing of the cover rock allows oil to migrate vertically to a much shallower depth. Wherever differential pressures exist, and permeable openings such as fractures provide a path, petroleum will move.

The composition and character of the oil ultimately formed is controlled by all phases of its origin and migration. The organic structures from which the oil originated, the selectivity of the migration phase in picking up only part of the hydrocarbons from the source rock, and the alteration of the oil within the reservoir by subsurface water and ground movements, are believed to play an important part in determining the composition of petroleum. See PETROLEUM; PETROLEUM GEOLOGY; SEDIMENTATION (GEOLOGY). [J.M.HU.]

**Bibliography:** G. D. Hobson, *Some Fundamentals of Petroleum Geology*, 1954; A. J. Levorsen, *Geology of Petroleum*, 1954; F. M. Van Tuyl, B. H. Parker, and W. W. Skeeters, The migration and accumulation of petroleum and natural gas, *Quart. Colo. School Mines*, 40:1-112, 1945; L. G. Weeks (ed.), *Habitat of Oil*, 1958.

## Petroleum engineering

The application of almost all types of engineering to the drilling for and production of oil, gas, and liquefiable hydrocarbons. It does not involve refining of liquid hydrocarbons nor long-distance transportation, which are respectively covered by refinery engineering and pipeline and marine engineering. See PETROLEUM PROCESSING; PIPELINE.

Petroleum engineering applies civil, mechanical, and chemical engineering processes, as well as thermodynamics and hydrodynamics, to the problems of producing hydrocarbons. Examples of various problems in petroleum engineering are discussed, but the various problems are too numerous and complex to permit complete consideration.

The petroleum engineer uses civil engineering in surveying, planning surface transportation systems, and in the design and selection of oil field equipment. Mechanical engineering is involved in the operation and utilization of gas and diesel engines, electrical motors, and rarely steam engines and their related equipment. Thermodynamics is the basis of understanding underground movements of hydrocarbons, their phase relations, and the recovery of liquefiable hydrocarbons from gases. Hydrodynamics is involved in fluid flow in the pay formation, vertically in tubing or casing, and in surface gathering and distribution systems. Chemical engineering is entailed in the proper design of the drilling mud, the protection of equipment against corrosion, and the treating of both oil and gas after production. See OIL AND GAS FIELD DEVELOPMENT.

**Drilling.** In drilling a wildcat well, a petroleum engineer must survey the location to make certain that the well is drilled at the desired distance from lease lines and wells on the property covered by the engineer's lease. He must provide access roads, fuel and water supplies, suitable foundations for the drilling equipment, and mud pits for storage of drilling mud, for settling of cuttings from the mud, and for treating the mud as may be necessary.

More than 80% of modern wells and all deep wells are drilled by the rotary process. A drilling

bit is turned at the bottom of the hole; mud pumped from the surface through the drill pipe cleans and cools the bit, flushes to the surface the rock cuttings that have been torn loose by the bit, and provides pressure to prevent collapse of the sides of the holes before casing is inserted. The mud must be heavy enough to prevent flow of fluids into the hole during the drilling operation. In soft or unconsolidated formations, care must be taken to prevent the hydrostatic pressure of the mud from forcing the mud into the surrounding formation. Excessive mud weight may result in loss of mud to such an extent that gas, oil, or water may break into the hole and, erupting at the surface, destroy the drilling equipment and tear out a crater at the surface, with resulting loss of hydrocarbons and damage to nearby properties. *See BORING AND DRILLING, MINERAL; OIL AND GAS WELLS.*

In some areas mud is formed by simply pumping water down the drill pipe and mixing this water with the soft clays which have been cut loose by the bit. But drilling mud is often an expensive and carefully engineered fluid produced by additions to the natural drilling mud. The cost of drilling mud varies from zero under most favorable conditions to costs exceeding \$10 per barrel of fluid. It is not uncommon for mud costs to exceed 20% of the total cost of drilling under difficult conditions. Careful chemical and physical testing may be required at short intervals to make certain that the mud has the desired characteristics and to plan for varying those characteristics as the bit penetrates different formations in depth. *See GEOPHYSICAL EXPLORATION; PROSPECTING; WELL LOGGING (MINERAL).*

The well having been drilled to a specified depth or to the point where commercial deposits of hydrocarbons are expected, tested, or proved, the hole is protected by "running" (inserting) a string of casing. The engineer must select casing of proper strength, diameter, and wall thickness so that the string can be lowered into the well without parting under the strain of its own weight: it must also have adequate collapsing strength so that it will not be crushed by outside pressures when fluid is removed from the casing. Commonly, important savings result from varying the strength and wall thickness in different parts of a casing or tubing string. Maximum tensile strength is required at the top of the string to sustain most or all of the entire weight. Maximum collapse strength is required at the bottom of the string where external pressures are at a maximum.

The casing is normally protected by pumping neat cement through it and permitting this cement to rise along the outside of the casing toward the ground surface. Cement quality must be chosen with full consideration of underground temperatures and fluids to allow ample time for pumping the cement to the desired position before the cement sets. If the cement does not rise to the desired point outside the casing, the casing must be perforated at one or more points and additional cement pumped through the perforations to form a cement sheath around the casing. When cementing casing, a plug

is placed in the casing at the top of the fluid cement. Mud pumped above the plug forces the plug to a point near the bottom of the hole so that little cement is left in the casing when the pumping operation ceases.

Frequently, and normally in deep wells, more than one string of casing is required to complete a well. For example, the normal hydrostatic pressure of fluids existing in the penetrated formations is 0.465 psi per foot of depth. In some areas, notably the Gulf Coast of Texas and Louisiana, certain formations contain fluids under pressures approximating the weight of the overburden so that the hydrostatic pressures in these abnormal zones may be 1 psi per foot of depth. Counterbalancing such a pressure requires mud weighing 18 lb/gal. If such heavy mud is used in formations of normal hydrostatic pressure, the mud may be forced into the surrounding formations and fail to return to the surface; the well may then be in danger because cuttings which should have been removed from the well by the circulating mud may settle and allow pipe to stick. Accordingly, the well is drilled with 10–12-lb mud to the top of the formation where abnormal pressure is anticipated, and casing will be landed and cemented at that point. The mud weight is then greatly increased for the zone of abnormal pressure. The petroleum engineer must design his mud and his casing strings to protect against such varying conditions.

Surveys are run during drilling to secure data on the straightness of the hole and the temperatures encountered. Unsurveyed wells have wandered (deviated) as far as 2000 ft distant from a vertical line through the well location. Electrical, neutron, and gamma-ray surveys determine the porosity of the formations penetrated and the character of the fluids in any porous rocks. Most of these surveys are made by specialists, but the capable petroleum engineer is prepared to interpret the surveys and to adapt his operations to the conditions shown by such surveys.

**Gas-oil separation problems.** Hydrocarbons occur underground under pressures and temperatures normally far in excess of surface conditions. The engineer must plan to separate the oil from the gas at the surface, direct the oil after measurement to pipelines, and, where necessary, transport the gas to gasoline plants where liquefiable hydrocarbons are removed from the gas; the gas is there dehydrated, sometimes treated to remove sulfur compounds, and residue gas is made available for sale to gas pipelines or for use in repressuring or recycling a producing formation. *See OIL AND GAS FIELD EXPLOITATION; PETROLEUM; PETROLEUM GEOLOGY.*

Under high pressures and temperatures all the gas may be dissolved in oil so that the formation contains only a single liquid-hydrocarbon phase. Conversely, if the volume of heavy hydrocarbons is relatively small, all of the oil may be dissolved in the gas so that the single hydrocarbon phase underground is a gaseous phase. If pressures are reduced, some hydrocarbon liquids condense from the

gas, form liquids in the pay formation, and are held by capillary attraction and thus lost. If a gas pool contains much liquid in gas solution and if this pool is produced as an ordinary gas pool, the loss of condensate (liquefiable hydrocarbons) in the pay may exceed 50% of the liquefiable content, a serious economic loss to the producer. Further, these liquids, by blocking minute connections between the pore spaces, may seriously impede the flow of gas into the well and thus extend the productive life of the field as much as 100%. Consequently, ultimate gas recovery may be reduced as much as 25%. These unfortunate results may be prevented by gradually withdrawing the gas, stripping it of liquefiable hydrocarbons, and reinjecting the dry gas under high pressure into the pay. This operation, known as recycling, is expensive because the returned gas must be compressed to a higher pressure than the existing pressure in the pay. High-pressure gathering, compressing, and distribution facilities are required. Extremely accurate engineering must be applied to determine in advance whether the cost of recycling will be greater or less than the value of the recovered liquids, and also whether the benefit of reduction in operating time, when the pool is operated without recycling, exceeds the net profit resulting from stripping the liquids from the gas before the available gas is marketed. See PETROLEUM RESERVOIR ENGINEERING; PETROLEUM SECONDARY RECOVERY.

**Recovery problems.** Oil pools produce under four mechanisms: gas expansion, gas-cap drive, water drive, and gravity drainage. Commonly, two or more mechanisms operate in a single pool. The petroleum engineer must be alert to take full advantage of any possible water drive or gas cap drive which will permit recovering a larger percentage of oil in the pay formation than can be obtained under gas-expansion drive.

Frequently, by exercising care in the location of wells and rate of production from individual wells, recovery from a pool may be double that which would have been secured by gas expansion or gravity drainage. Sometimes gas, water, or even air may be injected into a pool to displace and force oil into the producing wells. Injection of extraneous material and resulting increase in oil production is referred to as secondary recovery.

Water drives are the most efficient form of secondary recovery and may be expected to produce twice as much oil as may be secured by gas injection. Efforts to improve secondary recovery by water drive involve the use of materials which will reduce the surface tension of oil in the pay and so reduce the effect of capillary attraction in holding oil in the pay. Additives placed in the water may increase its surface tension to the point where it will displace oil. Other additives may reduce the surface tension of the oil and promote its displacement. Additives are expensive and their use is not always justified. Promising experiments include the injection of light hydrocarbons such as propane and butane ahead of the water. These light liquids, combining with the oil, reduce its viscosity, and in-

duce more complete flushing from the pay. Natural gas, carbon dioxide, or air are sometimes introduced into the water drive to assist in moving oil from the pay. The petroleum engineer must determine whether secondary recovery operations are justified; if so, he must select the method and the additives, if any, which will give the largest economic return from the money invested in the secondary recovery work.

**Corrosion.** Throughout the productive life of a field, the engineer faces the problem of corrosion. Carbon dioxide in gas, hydrogen sulfide in gas and oil, and salt water all corrode casing, tubing, sucker rods, tanks, and other equipment. These may be combated under varying conditions by selection of special steels containing varying percentages of nickel, chromium, and other alloying metals. Corrosion combatants, such as an alkali or formaldehyde, are frequently used to neutralize hydrogen sulfide and carbon dioxide. Some success has been obtained by coating sucker rods and internally coating tubing, surface lines, and equipment with plastics. In general, paints have not been successful in protecting equipment against corrosive fluids.

**Oil-water separation.** The engineer also applies chemical engineering in treating oil to separate water from oil emulsions. The treating may consist of merely heating the oil to facilitate the separation of water from oil. Sometimes the well fluid is filtered through hay, straw, or excelsior, saturated with oil, which rejects the water and permits clean oil to flow through the filter. A wide range of chemicals is used, depending on the chemical composition of the oil, to break down the emulsions. In some areas, electrostatic separation of oil from water is achieved by passing the fluid between insulated, electrically charged plates. Combinations of all these processes are successful in some areas. Conversely, some oils must be washed with water to remove sodium chloride and other salts from suspension. Again, the successful engineer must select the process best suited to the conditions existing in the field from which production is obtained.

**Production forecasts.** The petroleum engineer must predict the production of gas and of oil both in volume and time. The number of wells to be drilled, the rate of production, and the type of equipment to be used will vary, depending on the prediction as to life of the field and the volumes of oil, gas, and water to be produced. Predictions are made most roughly by a volumetric calculation which attempts to determine the volume of porous rock which is saturated with oil, the percentage of porosity in the rock, the percentage of water saturation in the pore spaces, the shrinkage of the oil resulting from release of gas, and the percentage recovery of the available oil. Necessary corrections are made for the differences between temperature and pressure in the formations and under standard conditions. Corresponding estimates are made for gas production. When a pool is first discovered, the only reserve estimate which can be made is based on volumetric calculations which are necessarily

inaccurate because the engineer does not have complete data on the various parameters. See PROSPECTING, PETROLEUM.

After a pool has some producing history, the engineer may estimate the ultimate production on the basis of curves comparing production against the declining pressure or decreasing rate of production. In gas pools, the accurately determined volume of gas produced per pound of pressure decline usually provides a good estimate of future production. However, an unanticipated active water drive in the pool may cause an excessive estimate. A material balance equation gives the most accurate measure of available reserves, provided necessary data are available. Estimates of productive life depend upon estimates of ultimate production and the estimated rate of production, which rate normally decreases as pool pressures diminish. Production rate should never exceed the maximum efficient rate of production, MER, which is the highest rate of production which can be maintained without underground waste. MER is frequently decreased by conservation regulations, by limited marketing facilities, and by delays in effecting unitization or other arrangement necessary for secondary recovery operations.

Predictions of reserves and of producing life are fundamental to the proper management of the field, the financing of the production operations, and the program for exploration in search of other productive fields. [R.S.K.]

**Bibliography:** R. L. Huntington, *Natural Gas and Natural Gasoline*, 1950; M. Muskat, *Physical Principles of Oil Production*, 1949; S. J. Pirson, *Oil Reservoir Engineering*, 2d ed., 1958.

## **Petroleum geology**

The application of geologic principles in the discovery and development of oil and gas pools. The geology of petroleum includes the origin, migration, and accumulation of petroleum; the structural and stratigraphic relations of oil and gas pools; the lithologic characteristics of formations and producing horizons; and the use of index or guide fossils in correlating horizons.

Geological aspects treated here include the occurrence of petroleum, the character of reservoir rocks, typical reservoir traps, and the general nature of reservoir fluids; for further treatment of the physical and chemical properties of petroleum, the origin and migration of petroleum, reservoir and production mechanics, and geological and geophysical methods of exploration, see PETROLEUM; PETROLEUM (ORIGIN); PETROLEUM ENGINEERING; PROSPECTING, PETROLEUM; see also MICROPALaeontology; PALYNOLOGY.

### **OCCURRENCE OF PETROLEUM**

Petroleum deposits may be classified as surface occurrences and subsurface occurrences.

**Surface occurrences.** These occurrences may be thought of as currently active or "live" occurrences, such as seepages, springs, exudates, and

mud volcanoes and mud flows. Others may be termed fossil or "dead" occurrences, such as bitumen-impregnated sediments, inspissated deposits, and dike and vein fillings of solid bitumens. Another surface occurrence of petroleum is oil shale, a borderline material between the petroleum hydrocarbons and the coal family.

**Seepages, springs, and exudates.** Petroleum that exudes in any of these forms may reach the surface along fractures, joints, fault planes, unconformities, or bedding planes, or through the connected porous openings of the rocks. Most seepages (or springs) are formed by the slow escape of petroleum from fairly large accumulations that have been brought close to the surface and into the zone of fracturing by erosion, or that have been tapped by faults and fractures. Almost without exception, seeps are at topographically low spots where water has also accumulated. Oil, which is lighter, rises to the surface of the water, covering it with an iridescent film. Many pools and producing regions have been discovered by drilling near seepages.

Exudates of asphaltic oils issuing at the surface are likely to be changed to asphalt, partly by the escape of volatile fractions, but mainly by chemical changes such as combination with oxygen or sulfur. The asphalt is black in color and varies in consistency from a sticky liquid to a substance hard enough to walk on. Outcrops of asphaltic oils are sometimes marked by small pools (less than 100 ft across) which have collected on the surface of the ground. Many of these pools contain the bones of animals that have been caught in the sticky material. Asphalt found at the surface has nearly always seeped up from the bedrock in the vicinity, but there are a few instances where asphalt has been carried by water. Asphalt of natural origin has been found floating in the Gulf of Mexico from Padre Island to Matagorda Peninsula, Texas.

**Mud volcanoes and mud flows.** These are high-pressure gas seepages that carry with them water, mud, sand, fragments of rock, and occasionally oil. Mud volcanoes are usually confined to regions underlain by incompetent softer shales, boulder and submarine landslide deposits, clays, sands, and unconsolidated sediments. The surface of a mud volcano is often a conical mound or hill, with an opening or crater at the top through which issue mud and water which is usually salty. Only the type which emits gas, with or without oil, in addition to the mud and water, should be considered a surface indication of oil or gas. Mud volcanoes occur chiefly in areas of Cenozoic rocks that have been strongly deformed. See MUD VOLCANO.

**Solid and semisolid deposits.** Tar, asphalt, wax, and hard brittle bitumen (any of the flammable, viscid, liquid or solid hydrocarbon mixtures soluble in carbon disulfide) are popularly regarded as solid, although strictly speaking, some of them are highly viscous liquids. Outcrops of solid petroleum are found in the form of disseminated deposits and as veins or dike-like deposits filling cracks and fissures.

Disseminated deposits are sediments containing petroleum in the form of asphalts, bitumen, pitch, or thick heavy oil, disseminated through the pore spaces of rock either as a matrix or as the bonding material. They are commonly called bituminous sands or bituminous limestones, depending on the nature of the host rock. Two different types of disseminated occurrences are found, inspissated deposits and primary mixtures of rock and bitumen.

Inspissated or dried-up deposits are in situ (in place) and were probably once a pool in liquid and gaseous form. They now consist of only the more resistant and heavier residues, the lighter fractions having been lost. An inspissated deposit may be thought of as a fossil oil field. As erosion gradually removed the overburden, bringing the surface closer to the pool, the decreased pressure permitted gases and lighter oil fractions to come out of solution and expand, leaving the heavier hydrocarbon fractions behind. As the pool approached the zone of weathering, the opening of incipient fractures allowed the gases to escape more readily. Oxidizing agents aided in solidifying the heavier oils that remained behind.

Primary mixtures are those in which the sediments were mixed with the oil, asphalt, or tar during their deposition, the whole deposit having later been buried by younger sediments and then exposed by erosion. The Athabasca oil sands in Alberta, Canada, are thought by many to be such a deposit. One theory is that oil seeped up from the underlying and then outcropping Devonian organic limestone and was redeposited during Cretaceous time together with Cretaceous sands in lagoons and barred basins along the shore. These oil sands may be considered a primary disseminated deposit in which the sand was deposited in or with the oil. *See OIL SAND.*

Dike and vein fillings may be regarded as fossil or dead seepages from which the gaseous and liquid fractions have been removed, leaving only the solid residues behind. In inspissated deposits, the separation of the lighter constituents occurred in place in the rock. In the primary deposits the separation of the gas from the liquid took place before the contemporaneous deposition of the oil and asphalt with the enclosing sediments. In the solid vein and dike fillings the loss of the gaseous and liquid fractions probably occurred while the petroleum was filling the opening.

Oil shale is rock which yields abundant oil on distillation, generally about 5–10 gal or more to the ton. It is a borderline class of hydrocarbon material, having some of the properties of petroleum and some of coal. Oil shale consists of solid hydrocarbons in the natural state. But they do not decompose into gaseous and liquid petroleum hydrocarbons until they are heated to temperatures of 350°C or more. The solid organic substances in oil shale which yield oil when heated were once called kerogen, but so many different meanings have been given to the word that it is not in good scientific standing today. *See OIL SHALE.*

Asphalt is a dark-colored, plastic to fairly hard substance, easily fusible and soluble in carbon disulfide. It occurs in nature, but it is also obtained as the residue from the refining of certain petroleum; then it is known as artificial asphalt. Asphalt melts between 150 and 200°F. It may occur as seepages, surface accumulations, and impregnations. Asphalt may also occur in large lakes. One of the best examples of such a lake is the Rancho La Brea deposit in Los Angeles.

Asphaltites are harder solid hydrocarbons which differ from asphalt in being strictly of an intrusive nature. They are found in veins or dikes cutting across the sediments. Asphaltites are fusible, but melt at somewhat higher temperatures and are harder and heavier than the asphalts. *See ASPHALT AND ASPHALTITE.*

Naturally occurring mineral waxes are solid hydrocarbons believed to result from the drying-out of a paraffin-base oil. One example is ozokerite, a plastic waxlike paraffin vein material which can be found in Utah and near Boryslaw, Poland. *See HATCHETTITE; OZOKERITE.*

**Subsurface occurrence.** Underground occurrences of petroleum may be classified as pools, fields, and provinces.

**Pools.** Underground accumulations of petroleum characterized by a single and separate natural reservoir (usually a porous sandstone or limestone) and a single natural pressure system are called pools. The production of petroleum from one part of a pool affects the reservoir pressure throughout its extent. A pool is bounded by geologic barriers in all directions, such as rock structure, impermeable strata, and water in the formations, so that the pool is effectively separated from any other pools that may be present in the same district or on the same geologic structure.

**Fields.** An oil field may be a single pool, or it may consist of two or more pools, all on or related to the same geologic structure. Where more than one pool is present in the same field, the different pools are separated from one another. The different pools may occur at several horizons of different geologic age, separated by impervious formations, and they may partially or completely overlap one another horizontally, or they may not overlap at all. Geologic features that are likely to form fields are salt plugs, anticlinally folded multiple sands, and complex combinations of faulting, folding, and stratigraphic combinations.

**Provinces.** A province is a region in which a number of oil and gas pools and fields occur in a similar or related geological environment. The term is used to indicate the larger producing regions, such as the Texas Panhandle and the Mid-Continent regions of the United States.

#### RESERVOIR ROCKS

Reservoir rocks are rocks with sufficient porosity and permeability to allow oil and gas to accumulate and be produced in commercial quantities. There are three requisites for a reservoir rock:

(1) it must be porous, that is, have enough room to store a commercial quantity or volume of hydrocarbons; (2) it must have permeability so that the contained oil or gas will discharge readily when the reservoir is penetrated by a well; and (3) there must be a trap which prevents escape of the oil and gas until they are released by the drill bit. Any rock with these characteristics may become a reservoir for migrating hydrocarbons.

The reservoir character of a rock may be an original feature (intergranular porosity of sandstones), or a secondary feature resulting from chemical changes (solution porosity of limestones), or it may be the result of physical changes (fracturing of any brittle-type rock).

**Types of reservoir rocks.** Reservoir rocks may be classified as fragmental or clastic (broken); chemical and biochemical; or miscellaneous. They may also be classified as marine and nonmarine reservoir rocks.

*Fragmental type.* Some reservoir rocks are aggregates of particles, that is, fragments of minerals or older rocks. They are also called clastic or detrital rocks because they consist of mineral and rock particles derived from eroded areas. The constituent particles of fragmental rocks may range in size from colloidal particles up to pebbles and boulders. The most common of the fragmental reservoir rocks are siliceous—sandstones, limestones, conglomerates, arkoses, graywackes, and siltstones. Many, however, are carbonate rocks, such as oolitic rocks and coquinas, which are made up of oolites and shell fragments that have been only slightly cemented or recrystallized.

Some sandstone reservoir rocks consist either entirely or in part of loose, uncemented sand grains. The grains are brought to the surface in large quantities along with oil during production. The sand grains in most sandstones, however, are held together by various kinds of cementing material, mostly carbonates, silica, or clays. Some of the cementing materials may be primary, having been deposited along with the sand grains and then precipitated chemically around and between them. Other cementing material may be secondary, having been precipitated from water solutions that entered the formation after it was deposited.

Clastic limestones and dolomites consist of grains of calcite and dolomite that have been transported and deposited just as are grains of quartz. The carbonate grains are made up largely of shells, shell fragments, coquina, and oolites. Rocks thus formed are always more or less recemented with recrystallized calcite and may resemble a chemically deposited limestone or dolomite. Carbonate rocks formed this way are usually good reservoirs for oil because of their porosity.

*Chemical types.* These rocks are made up chiefly of chemical or biochemical precipitates. They are composed of mineral matter that was precipitated at the place where the rocks were formed (in contrast to the transported grains in clastic carbonates). The most important chemical reservoir

rocks are carbonate sediments, mostly limestones and dolomites. Some chemically precipitated rocks consist entirely or almost entirely of silica in the form of chert, novaculite, or orthoquartzite, but in some of these there has been a certain amount of secondary cementation with silica. Such rocks are quite common, but compared with the carbonate rocks they provide few reservoirs. The porosity of this type of rock is largely the result of solution which involves the leaching away of portions of the rock by percolating ground waters.

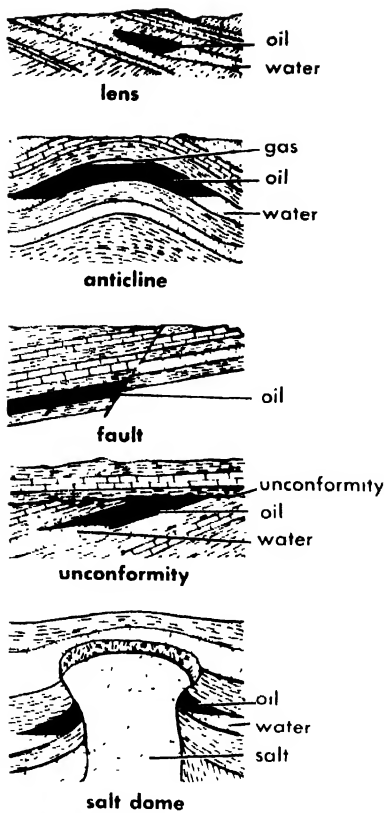
*Miscellaneous types.* Other reservoir rocks include igneous and metamorphic rocks and mixtures of both. Any porous and permeable igneous rock in close association with sedimentary rock may become a reservoir rock when saturated by oil derived from the sediments. Igneous and metamorphic rocks are only a minor source of oil and gas because, generally, they are not permeable enough, and when they are, they are not often associated with suitable source rocks and a good trap for oil and gas. Porosity and permeability of igneous and metamorphic rocks may result from fracturing or from weathering at the surface prior to subsequent burial.

*Marine and nonmarine types.* A distinction may be made between reservoir rocks which were deposited in ancient seas and those of continental origin. Until recently most petroleum was believed to occur in rocks deposited under marine conditions. Consequently there was little exploration of nonmarine reservoirs. However, petroleum has been found in sediments of nonmarine origin, such as those in the Uinta Basin of Utah, which consist of fresh-water lacustrine marls, limestones, and siltstones. The occurrence of oil in nonmarine sediments is sometimes explained as the result of migration of oil along faults, fractures, or bedding planes from adjacent marine sediments. Further study of nonmarine reservoir rocks will provide important information on the occurrence and source of oil deposits.

**Properties of reservoir rocks.** The porosity and permeability of reservoir rocks, as well as the nature of the traps, are all factors which regulate the accumulation of petroleum. Porosity is the total space in the rock (pores, voids, interstices) not occupied by solid material. It is expressed as a percentage. Factors which influence porosity are the size of the rock particles, arrangement, sorting, shape, cementing material, and the connate water content. Most oil-producing rocks have porosities above 10% and thicknesses greater than 10 ft. A rock with lower porosity may prove commercially exploitable if the thickness is great and a thinner rock may be developed successfully if the porosity is unusually large.

Total pore space is not the sole determinant of a petroleum reservoir. A reservoir must also have permeability; that is, fluids must flow through it with relative ease. Pumice, for instance, has a large amount of pore space, but the pores are not connected. Therefore it does not have permeability





Typical petroleum traps.

because it cannot discharge its fluid content. For a rock to be permeable, it must have porosity, inter-connecting pores, and pores of supercapillary size. Permeability is measured in Darcy units.

#### RESERVOIR TRAPS

Reservoir traps close the reservoir so that the accumulated oil or gas cannot escape. The upper boundary of a reservoir trap is called the roof or cap rock; the lower boundary is called the oil-water contact or table.

Roof or cap rock is an impervious layer of rock forming the roof of an oil trap. The connecting pores in the reservoir rock, which are individually minute, are as a rule saturated with water. Since oil and gas are lighter than water, the petroleum rises through the water until it is stopped by the roof rock. If the roof rock is concave (domed, arched, folded, peaked, or roof-shaped) it acts as a trap, keeping the oil and gas from escaping laterally.

Oil-water contact or table is the lower boundary of the reservoir. It usually consists either totally or in part of the water that normally fills the pores of the reservoir rock. The water supports the pool of oil and gas, and the pressure of the water forces the petroleum upward against the bounding surfaces of the trap, holding it in place. The simplest and most common way for a permeable underground formation to become a trap is to be folded into an anticline. An anticline is a fold (resembling an inverted soupbowl) that has upward convexity. It can be open or closed. However, only

those anticlines which have effective closure with some horizontal extent are productive of oil or gas.

**Geologic structure.** The anticlinal theory is the most successful of all the theories of petroleum geology. It has been estimated that fully 80% of the oil in the 236 major oil fields of the world is in anticlines. The fact that oil and gas commonly occur on anticlinal axes was first noted by W. E. Logan in 1842. He observed that oil seeps occurred in the vicinity of anticlinal axes near the mouth of the St. Lawrence River. Although the term anticlinal theory has fallen into disuse, the fundamental principle on which it is based is still valid—oil and gas accumulate at the greatest possible height within the reservoir. It is recognized today that other factors control the accumulation of oil in many pools.

**Classification of traps.** Three basic types of traps generally are recognized: structural traps, stratigraphic traps, and combination traps.

**Structural traps.** A trap whose upper boundary has been made concave by some local deformation such as folding or faulting (or both) of the reservoir rock is known as a structural trap. The edges of a pool occurring in a structural trap are determined by the intersection of the underlying water table with the enclosing roof or cap rock. Structural traps include closed anticlines or domes, faulted anticlines with closure, closure against faults, anticlines on downdip sides of faults, and oil and gas accumulations in fractures produced by structural deformation.

**Stratigraphic traps.** Also known as varying permeability traps, stratigraphic traps are those in which the chief trap-making element is some variation in the stratigraphy or lithology or both of the reservoir rock. These include facies change, variable local porosity and permeability, and any up-structure termination of the reservoir rock. Stratigraphic traps include sandstone lenses, channels, bars, and reefs; porosity lenses; and reservoirs in permeable organic solids (coal, oil shale). Some of the most common stratigraphic traps are strand-line pools, shoestring sand traps, biostromes, and bioherms. See FACIES (GEOLOGY).

Strand-line pools are regional facies changes from permeable to impermeable rocks which determine the location of the edges of an oil pool. They are so called when associated with shore phenomena. See STRAND LINE.

Shoestring sand traps are long, narrow, sand deposits which may be considered to be sand lenses of a special type. They may be one-half or three-quarters of a mile wide and to many miles in length. Except at their terminal ends they are completely surrounded by impervious shales and clays. Some sand traps of this nature are believed to be channel fillings and others offshore sand bars.

Two general classes of primary stratigraphic traps occur in rocks of chemical origin, almost all of them carbonate rocks. These are biostromes and bioherms. Biostromes are nearly tabular, porous lenses, either lithofacies or biofacies, enclosed or terminated by normal impervious shales, lime-



stones, or dolomites. Bioherms, or organic reefs, are porous, domelike, moundlike, or otherwise circumscribed masses, built exclusively or mainly by sedimentary organisms such as corals, algae, brachiopods, mollusks, or crinoids, and enclosed in normal rock of different lithologic character. See BIOHERM; BIOSTROME.

**Combination traps.** These traps result from both structural and stratigraphic conditions. An example of such a trap is the salt dome. Salt domes are cylindrical or steeply conical masses of rock material which were forced to flow plastically under heavy pressure. These masses, called plugs or domes, originate at unknown depths and pierce the overlying sedimentary strata. Three kinds of traps are associated with salt plugs: cap rock, flanking sands, and supercap sands. Cap rock consists of calcite, gypsum, and anhydrite and occurs as a capping over the tops of the salt plugs. Flanking sands are strata abutting upon and cut off by the salt plug. Supercap sands are sandy strata that arch over the tops of the plugs in the form of structural domes. See SALT DOME.

#### RESERVOIR FLUIDS

Fluids fill the voids or pore spaces in all reservoir rocks. The fluid may be water, water and oil, water and gas, or a mixture of water, oil, and gas. The fluid content of a gas pool consists of water and gas; that of an oil pool consists of gas, oil, and water. There is almost an infinite variation in the composition, relative amounts, and properties of these fluids in various reservoirs.

The distribution of gas, oil, and water in the reservoir depends upon relative buoyancy, relative saturation of pore space with each fluid, capillary and displacement pressures, as well as the porosity, permeability, and composition of the reservoir rock. In traps that contain gas, oil, and water, the fluids take on a zonal character. Gas, being the lightest, fills the pores nearer the top of the trap. Below the gas there is a zone in which the pore-filling liquid is chiefly oil, and below that, water alone occurs, the contact being the oil-water table. Where there is gas but no oil, the gas is immediately underlain by water and the contact is the gas-water table. Interstitial water (adsorbed water or wetting water which lines the pore walls) is present throughout the reservoir, occupying from a few to about 50%, but generally between 10 and 30%, of the pore space. The amount of interstitial water in a petroleum reservoir is commonly measured in percentage of effective pore space and is known as the water saturation.

Oil-field waters are waters associated with oil and gas pools. They may be classified as meteoric waters, connate water, and mixed water. Most oil-field waters are saline, except at shallow depth. See OIL-FIELD WATERS.

Oil saturation is the amount of oil contained in a petroleum reservoir. It is measured as a percentage of the effective pore space.

Gas volume, or natural gas content of a petroleum reservoir may range from small quantities dissolved in oil up to 100% of petroleum content.

Natural gas may be classified as associated when it occurs with oil, and as nonassociated when it occurs alone. The natural gas in a reservoir may occur as free gas, as gas dissolved in oil, gas dissolved in water, or as liquefied gas. See NATURAL GAS; OIL AND GAS WELLS; PETROLEUM RESERVOIR ENGINEERING. [D.D.G.]

**Bibliography:** G. D. Hobson, *Some Fundamentals of Petroleum Geology*, 1954; C. G. Lalicker, *Principles of Petroleum Geology*, 1949; K. K. Landes, *Petroleum Geology*, 2d ed., 1959; A. I. Levorsen, *Geology of Petroleum*, 1954; W. L. Russell, *Principles of Petroleum Geology*, 1951; E. N. Tiratsoo, *Petroleum Geology*, 1952.

#### Petroleum microbiology

Those aspects of microbiological science and engineering of interest to the petroleum industry, including the role of microbes in petroleum formation, exploration, production, manufacturing, and storage.

**Petroleum formation.** Dead marine microorganisms comprise much of the sedimentary material from which petroleum is formed. Other bacteria, such as *Pseudomonas*, *Achromobacter*, *Desulfovibrio*, and *Flavobacterium* species, modify the organic nature of this material. The extent to which they actually convert organic sediments to petroleum hydrocarbons is uncertain, because this has not yet been demonstrated in the laboratory.

**Petroleum exploration.** Many microorganisms are able to employ hydrocarbons as a carbon source, converting them either to carbon dioxide and water or to intermediate organic compounds. The soil above a petroleum reservoir may contain gaseous emanations, such as methane and ethane, from the reservoir, and it is believed that these may be detectable by the concentrations of certain hydrocarbon-utilizing bacteria (*Methanomonas* species) in the soil, or by the growth of suitable cultures planted there. Such exploration techniques have been intensively investigated, but their success remains questionable.

**Petroleum production.** Microorganisms (*Crenothrix*, *Beggiatoa*, *Pseudomonas*) are frequent and costly contaminants of drilling fluids and of secondary recovery injection waters; and chemical agents, like formaldehyde and quaternary ammonium compounds, are necessary for their control. Bacterial corrosion of iron pipe, particularly buried pipe, by *Desulfovibrio* species is a major problem in the petroleum industry. It is generally treated by protective coatings or by cathodic protection. Bacterial deterioration of refined petroleum products in storage, of asphalt and asphalt coatings, and of oil emulsions used with cutting machinery are also industrial problems. [E.J.BE.]

**Bibliography:** E. Beerstecher, Jr., *Petroleum Microbiology*, 1954.

## Petroleum processing

The recovery and processing of various usable fractions from the complex crude oils. The usable fractions include gasoline, kerosine, diesel oil, fuel oil, asphalt, lubricating oils, and many others.

The petroleum refining industry is one of the largest manufacturing industries. Over \$1,000,000,000 are spent each year in the United States for capital equipment, maintenance and modernization. Almost \$800,000,000 were spent in the United States for new equipment and facilities in 1958.

The United States alone produced 3,950,000 barrels of gasoline per day (bbl/day) in 1958. In addition, the following products were produced in the quantities shown: middle distillate (including kerosine, diesel oil fuels, and others), 2,090,000 bbl/day; residual fuel oil (for heating purposes),

1,500,000 bbl/day; all other products (including waxes, lubricating oils, asphalt, coke, and others), 1,590,000 bbl/day.

Crude oil is a mixture of many different hydrocarbon compounds of the paraffin type (wax compounds) and of the naphthene type (asphalt compounds), making the chemistry of petroleum refining extremely complex. The refining processes can be grouped under three main headings: (1) separating the crude oils to recover the desired products; (2) breaking the remaining large chemical compounds into smaller chemical compounds by cracking; (3) building the desired chemical compounds by chemical reactions, such as polymerization, reforming, alkylation, and isomerization.

Refinery products, such as gasoline, kerosine, diesel oil, and others, are not pure chemical compounds but mixtures of chemical compounds. Some of the hydrocarbon compounds contained in gasoline are shown in Table 3 along with the individual specific gravities, molecular weights, and normal boiling points.

A simplified flow sheet of refinery operations is shown in Fig. 1. By means of distillation a typical crude oil may be separated quite easily into many fractions of raw products. Some of these are shown in Table 4.

A more complex flow sheet of a refinery for light oils is shown in Fig. 2. Here are included the cracking equipment, reforming equipment, extraction units, polymerization units, and other facilities. Figure 3 is a schematic diagram of a refinery for producing lubricating oils.

**Separating the crude oil.** There are two principal separating procedures not involving cracking—topping of crude oil, and lubricating oil processing. Both of these procedures include combinations of several operations, such as distillation, centrifuging, filtration, and treating processes.

**Topping, or distilling, the crude oil.** The crude oil is desalted and dehydrated, then passed through heaters where the temperature is raised to about 650°F, at which temperature all of the gas, gasoline, kerosine, and light fuel oil fractions are in the vapor phase. This vapor and liquid mixture enters a large distillation tower about one-third the distance up from the bottom as shown in Figs. 1 and 2.

Table 1. Refining capacities in the free world\*

Region or country	Refining capacity (crude oil charged)	
	Jan. 1, 1958	Jan. 1, 1959 (est.)
United States	9,800,000	9,998,000
Canada	810,000	1,010,000
Latin America	2,640,000	2,995,000
Europe	3,025,000	3,575,000
Middle East	1,265,000	1,540,000
Far East	1,090,000	1,150,000
Africa	100,000	125,000
	18,730,000	20,393,000

SOURCE: *World Oil*, Feb. 15, 1958.

\* Crude oil charging capacities in barrels per day.

Table 2. Capacities of refining operations\*

Operation	United States	Free World (other than U.S.)
Thermal cracking	1,005,000	2,350,000
Thermal reforming	210,000	45,000
Catalytic cracking	4,150,000	1,025,000
Catalytic reforming	1,625,000	320,000
Asphalt production	425,000	80,000
Lubrication oil production	200,000	57,000

SOURCE: *Petroleum Refiner*, September, 1958.

\* In barrels per day, Jan. 1, 1958.

Accurate data is not available for the Communist World.

Table 3. Some chemical compounds found in gasoline\*

Name	Formula	Molecular weight	API gravity	Normal boiling point, °F	Research octane number
n-Pentane	C <sub>5</sub> H <sub>12</sub>	72	92.7	97	62
n-Hexane	C <sub>6</sub> H <sub>14</sub>	86	81.6	158	25
n-Heptane	C <sub>7</sub> H <sub>16</sub>	100	74.1	209	0
n-Octane	C <sub>8</sub> H <sub>18</sub>	114	68.7	260	-19
n-Nonane	C <sub>9</sub> H <sub>20</sub>	128	64.6	310	-34
n-Decane	C <sub>10</sub> H <sub>22</sub>	142	61.3	343	-53
n-Endecane	C <sub>11</sub> H <sub>24</sub>	156	58.0	387	-60

\* Only the straight-chain paraffin hydrocarbons are shown here to indicate the range. Actually the gasoline contains also branched-chain paraffins, alkenes, naphthenes, aromatics, and other compounds.

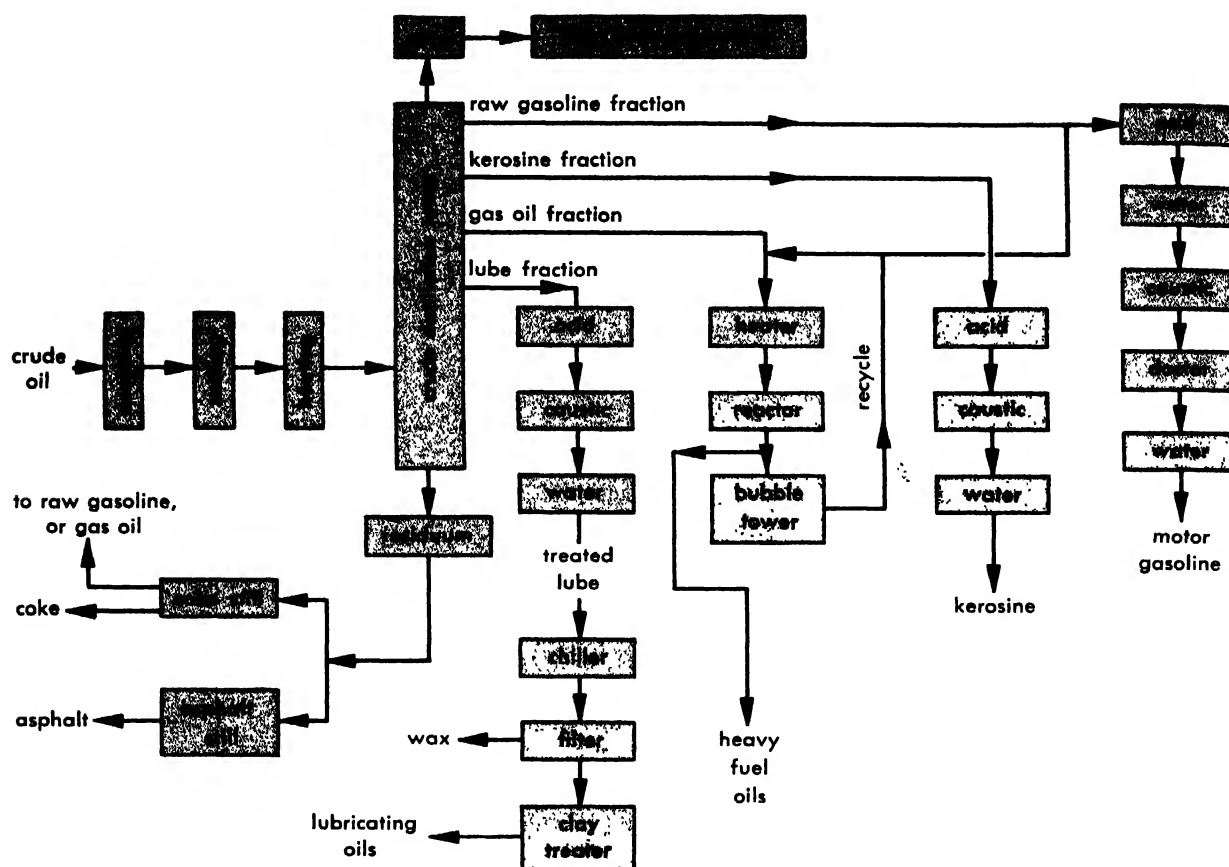


Fig. 1. Simplified flow sheet of crude oil refining.

Into the bottom of the tower about 1 2 lb of steam per gallon of crude oil are usually introduced to make the separation easier. From the top of the tower some gases are evolved and sent to units processing light ends. The next higher-boiling fraction is the gasoline, followed successively by the kerosine, the gas oil, the cracking stock, and the lubricating distillate. Below the feed entrance a fraction called the residuum is removed.

The temperature of the feed to the tower depends considerably upon the ultimate plans for the residual oil. If this residual oil is to be processed further for the manufacture of lubricating oils, the feed is not heated to so high a temperature.

Each of the streams from the distillation unit must be treated further before it can be sold. The gasoline fraction is treated, then blended with other stocks. Finally, certain chemicals called additives are added to the stream to improve its properties. See DISTILLATION; FRACTIONATING COLUMN.

Table 4. Some fractions obtained from crude oil

Fraction	Carbon atoms	Molecular weight	API gravity	Boiling range, °F
Gas	1-4	16-58	0.38-0.58	-259-31
Gasoline	5-12	72-170	58-62	31-400
Kerosine	10-16	156-226	40-46	356-525
Gas oil	15-22	212-294	34-38	500-700
Lube oil	19-35	268-492	24-30	640-875
Residuum	36-90	492-1262	8-18	875+

*Lubricating oil processing.* The most important property of lubricating oil is its viscosity. The lube fraction produced in the vacuum distillation column contains some hydrocarbons that give the oil a poor viscosity-temperature characteristic. In addition, the lube oil fraction has poor oxidation resistance and contains wax and other impurities which must be removed. Consequently, the lubricating oil fraction must be treated to remove or to reduce the concentrations of the following: free-carbon-forming material, low viscosity-index materials, wax, unstable compounds which may decompose to form asphaltic substances or coke, and chemicals that affect the color of the lube oil products.

The flow sheet shown in Fig. 3 describes a process for the production of lubricating oils. Not only the lubricating fractions but also a portion of the residuum fraction is used to make the lubricating oils. In this case, the residuum is treated with a solvent to remove the asphaltic material. The deasphalted residuum is further extracted along with the other lubricating oil fractions, dewaxed, acid-treated, clay-treated, blended with additives, and then sent to storage.

The solvents used for deasphalting include furfural, cresylic acid, phenol, sulfur dioxide, chlorex, nitrobenzene, propane, benzene and many others. Since the solvent must be removed from the oils after the extraction, elaborate distillation equipment is required. The oils are freed from the final

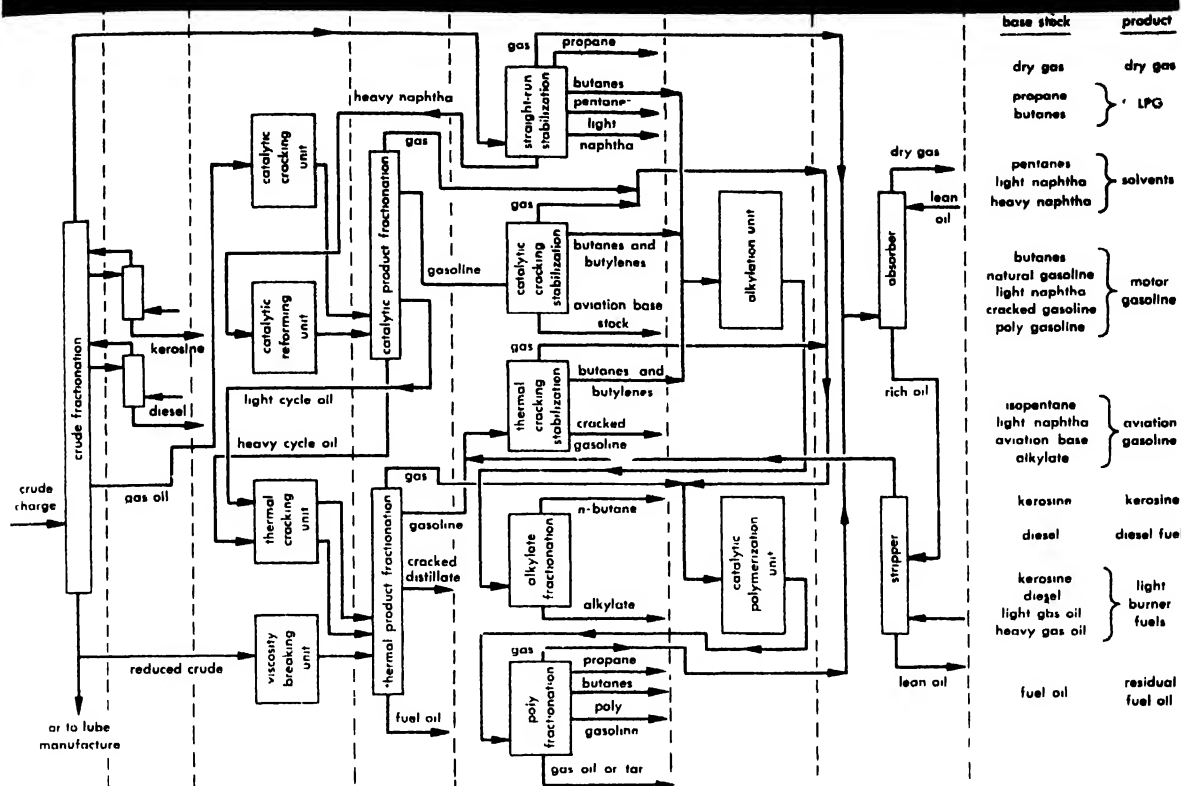


Fig. 2. Schematic diagram of a refinery for light oils (mainly gasolines, kerosine, and distillates).

traces of solvent by steam stripping or vacuum flashing. Quite often the solvent is more expensive than the oil being treated, so that from the standpoint of economy alone, all of the solvent must be recovered. More importantly, the solvent itself may have properties detrimental to the finished oil when present in trace amounts.

**Distillation.** All distillation processes are essentially the same. The factors to be considered for different types of distillation processes include the sensitivity of liquid with respect to heat, the specifications of the product, and the boiling range of the feed.

In topping or skimming procedures, the crude is heated to a certain temperature and fed to a distillation tower where the product fractions are removed at various heights along the column. Figure 1 includes a schematic diagram of such a separation.

**Stabilization** is the distillation process that removes the lighter hydrocarbons (usually the dissolved gaseous hydrocarbons) from the particular fraction being processed. Here the feed is heated and sent to a fractionation column, where gases are removed overhead and the stabilized product at the bottom. In natural gasoline stabilizers, 40–60 plates are required in the distillation column to remove the dissolved propane and lighter hydrocarbons. In the stabilization of pressure distillate the feed is heated to a much higher temperature since less propane and butane are present.

Steam distillation is used to increase the amount of distilled products obtainable at a fixed feed temperature. The feed stock is heated to approximately 550–660°F in the presence of a large amount of steam. The effect of steam is to reduce the boiling point by partial pressure effects. The boiling point of a material can be reduced either by reducing the total pressure or by adding an inert gas such that the same total pressure will be partially due to the inert gas.

Vacuum distillation is used for the redistillation of the pressure distillate, lube stock, topped crudes, and other fractions. Lubricating oil, for example, is thermally sensitive and partially decomposes if exposed to high temperatures. Therefore the distillation is done under a high vacuum to take advantage of the lower temperatures required at the lower operating pressures. Sometimes, high vacuum is not sufficient, and it is necessary to combine vacuum distillation with steam distillation in a combination unit. In this case, steam is added to the distillation column operating under the vacuum. The amount of steam required will vary, of course, but may be as high as 1 or 2 lb/gal of oil processed. The dry vacuum distillation processes have the advantages that smaller towers are required for a given throughput. In addition, smaller condensing equipment is required.

**Centrifugation.** The centrifuging process for dewaxing lube stocks is being replaced rapidly by solvent dewaxing processes. However, many cen-

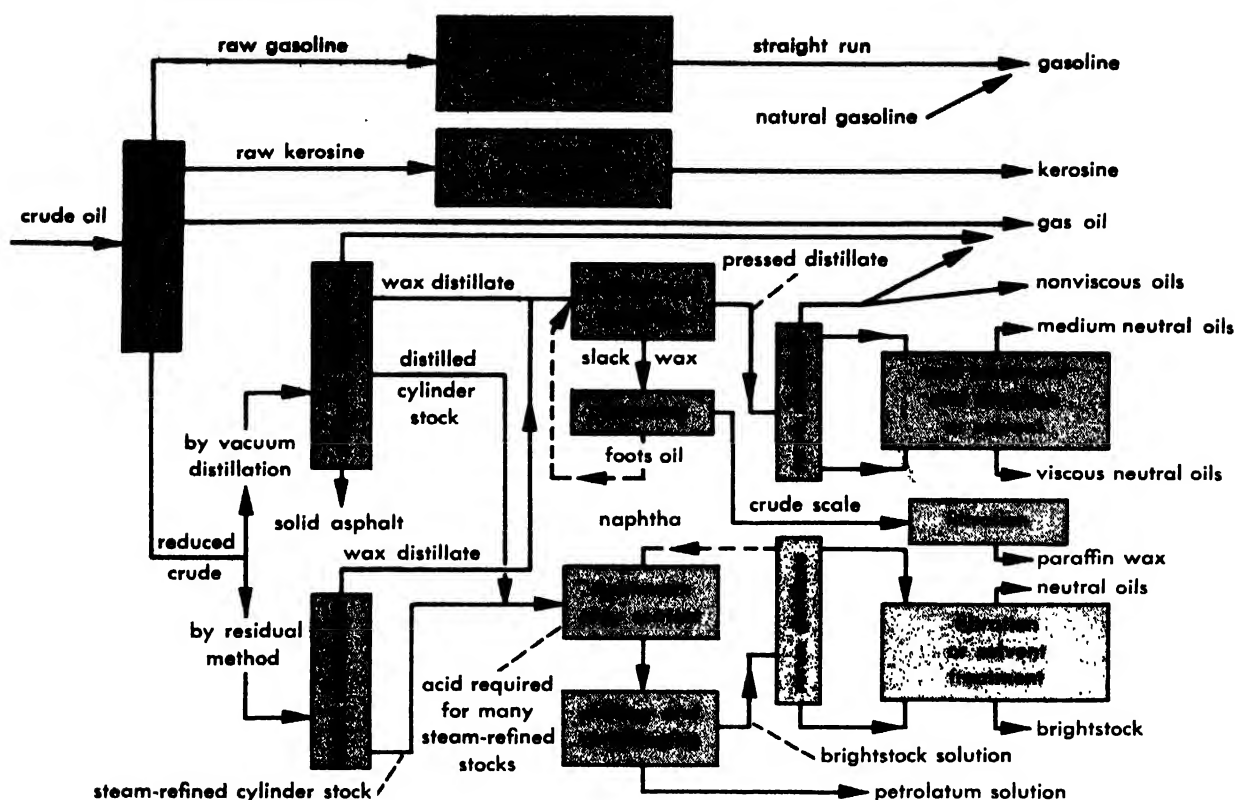


Fig. 3. Schematic diagram of a refinery for lubricating oils (from W. L. Nelson, *Petroleum Refinery Engineering*, 4th ed., McGraw-Hill, 1958).

trifuging processes are still in operation. In this process the lube stock is mixed with about 50–80% naphtha and chilled to some low temperature. If the pour point of the finished oil is to be 20°F then the stock is chilled to approximately –10°F; for a pour point of 0°F it is necessary to chill to –40°F. The cold solution is fed to a centrifuge where the wax is separated from the oil. The capacity of one of these centrifuges may be as high as 75–100 bbl/day of oil. The centrifuges operate at around 16,000 rpm and require approximately 1 kw of power. See CENTRIFUGATION.

**Filtration.** This is also an important operation in the refining of petroleum. Regular gravity-type settlers are used wherever possible, but occasionally the solids are too finely divided to settle. There are many types of filters used for the removal of finely divided clay from treated stocks in the clay-contact process. These filters are classified as filter presses, leaf filters, rotary filters, and others. See FILTRATION.

**Breaking the large molecules.** The major product from the refinery is the motor fuel (gasoline). Of course kerosine, diesel oils, jet fuels (mostly kerosine fraction), and others are extremely important also. However, each barrel of oil charged to the distillation tower has a given fraction of gasoline. This varies but on the average is not over 20% of the total volume of crude. If more gasoline than this 20% obtainable by distillation is desired, and it almost always is, it is necessary to resort to other

means than straight separation to get more gasoline. This can be done either by recombining the gaseous, or lighter, molecules (polymerization), or by breaking down the heavier molecules (cracking).

**Cracking.** Table 3 shows that gasoline molecules seldom contain more than 11–12 atoms of carbon. The crude oil, however, contains many molecules consisting of more than 50–60 atoms of carbon. The heavy naphtha fraction and the kerosine and gas oil fractions, for example, all contain large molecules compared with the gasoline fraction. In order to use these fractions for gasoline production, the long or large molecules must be broken into smaller ones of the gasoline type. This process is called cracking.

The cracking may be done either by thermal means (maintaining the heavy fractions at high temperatures) or by catalytic means. In thermal cracking the charge stocks are usually light and heavy gas oils, residual oils, or any of the topping column fractions heavier than the gasoline fraction. The resulting gasoline yields depend upon the composition of the charge stock but will range from 15 to 40% by volume of gasoline (100–400°F boiling range).

In catalytic cracking the fraction to be cracked is contacted with a catalyst under lower pressure conditions than in thermal cracking, although the temperatures are still almost as high. Catalytic cracking gives much better yields of gasoline, lower

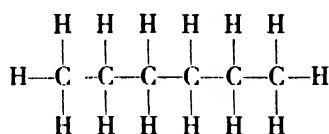
carbon formation, and a gasoline of much higher octane number.

About 80% of the cracking capacity in use in the United States today is of the catalytic type. See CRACKING.

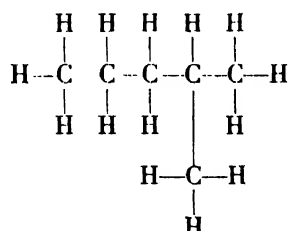
### Rebuilding the desired chemical compounds.

The saturated straight-chain paraffins shown in Table 3 have very low octane numbers. These compounds can be altered, however, by chemical reaction to yield a different kind of molecule with much higher octane characteristics. In general, the straight paraffin compounds have the lowest octane rating and the aromatic compounds (benzene family) have the highest. The olefins and the naphthenes have intermediate octane numbers.

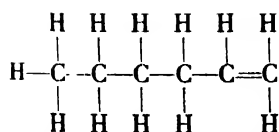
Below are some of the forms of a six-carbon-atom hydrocarbon and their research octane numbers (RON). All these forms, and many others, are found in the gasoline fraction. It is possible to convert hexane (RON = 24.8) into benzene which has an octane number of over 100.



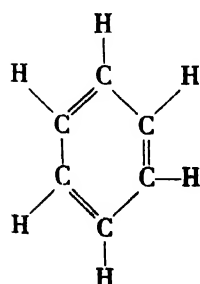
*n*-Hexane, C<sub>6</sub>H<sub>14</sub> (straight-chain paraffin) RON=24.8



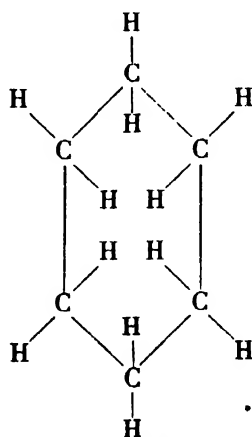
2-Methylpentane or isohexane, C<sub>6</sub>H<sub>14</sub> (branched paraffin or isoparaffin) RON=73



1-Hexene, C<sub>6</sub>H<sub>12</sub> (olefin or alkene) RON = 80



Benzene, C<sub>6</sub>H<sub>6</sub> (aromatic) RON = over 100



Cyclohexane, C<sub>6</sub>H<sub>12</sub> (naphthene) RON = 83

1. Hydrogenation is used mostly for producing saturated hydrocarbons from unsaturated ones. During World War II, this process was used for making isooctane from isooctene. More recently this process has been used almost exclusively for desulfurization processes.

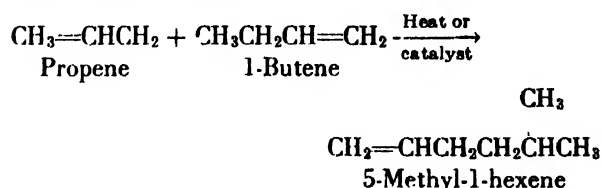
2. Dehydrogenation is the removal of hydrogen from a molecule. For example, 1-hexene may be made from *n*-hexane. This reaction often results in increased octane number.

3. Aromatization yields aromatic type hydrocarbons from other types, as benzene from hexane or cyclohexane. Aromatization and isomerization predominate in the reforming operation.

4. Cyclization is the transformation of a hydrocarbon of the chain type to one of the ring type; for example, making methylcyclohexane from *n*-heptane.

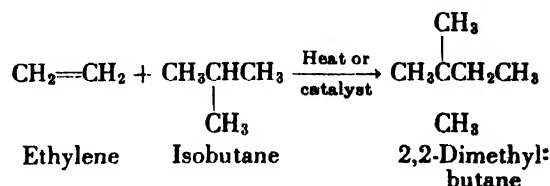
5. Isomerization is the rearrangement of the atoms in a molecule, such as *n*-hexane, to form an isomer, such as isohexane. In 1958, the isomerization capacity in the United States refining industry was about 30,000 bbl/day, but this promises to be a rapidly expanding process in the industry.

6. Polymerization involves two or more molecules in a building process. For example, propene and the butenes, which are present in the gases of thermal- or catalytic-cracking operations, are polymerized to form a larger liquid molecule with a high octane number:



At the present time, the catalytic polymerization capacity in the United States is over 140,000 bbl/day.

7. Alkylation also makes use of two or more molecules in the reaction. This process uses an isoparaffin, such as isobutane, and an olefin, such as ethylene, to yield a larger molecule with a high octane number, 2,2-dimethylbutane, in this case:



This reaction uses only one-half as many of the expensive olefin molecules as the polymerization process. In 1959 the alkylation capacity in the United States was over 360,000 bbl/day. This is also a fast-growing portion of the refining industry. See ALKYLATION; AROMATIZATION; HYDROGENATION; ISOMERIZATION; OCTANE NUMBER; REFORMING (PETROLEUM REFINING).

**Treating processes.** Both the crude oil and the petroleum products must, on occasion, be treated to remove undesirable impurities or to improve the

Among the many processes used for altering the chemical structure of the molecules are the following:

properties of the product. The important treating processes are desalting and dehydrating, sweetening and desulfurization, acid treatment, clay-contact adsorption treatment, vapor-phase treatment, and solvent treatment. Some of these processes are used both on the crude oil and on the products, whereas others are used on only one or the other.

**Desalting and dehydration of the crude oil.** The salt content of the crude oil which enters the refinery may be as high as 4 or 5%, and the water content may be much higher than the equilibrium amount because water is present as an emulsion.

Because of the high temperatures of the heater tubes, the introduction of the wet crude into the heaters would be dangerous. In addition, the salt would precipitate onto the tube walls, reducing the rate of heat transmission and thereby the efficiency of the heaters.

Many processes are available for the removal of both the salt and the water from the crude oil. These are grouped into four types as shown in Table 5.

The crude oil (containing the salt and the oil) is heated, an emulsion breaker is added, and the resultant mass is settled, or even filtered, to remove the salt and water phase from the oil phase.

**Sweetening and desulfurization.** Since the original crude oils contain some sulfur compounds, the resulting gasolines and other products also contain sulfur compounds, including hydrogen sulfide, mercaptans, sulfides, disulfides, and thiophenes.

The processes used to sweeten, or desulfurize, the products depend upon the type of sulfur compounds present and the specifications of the finished gasoline or other stocks.

Mercaptans are removed or converted into less undesirable disulfides in these ways:

1. Mercaptan removal: (a) Unisol process uses an alkaline solution of methyl alcohol, (b) Solutizer processes use sodium hydroxide along with minute amounts of sodium isobutyrate, and (c) Mercapsol process uses an alkaline solution of naphthenic acids and phenols. These are regenerative solution processes.

2. Mercaptan conversion (oxidation to disulfides): (a) lead sulfide doctor sweetening, (b) copper chloride-oxygen sweetening, (c) sodium hypochlorite sweetening, and (d) copper sweetening.

3. Hydrogen sulfide removal by regenerative solution processes using aqueous solutions of the fol-

lowing: (a) sodium hydroxide, (b) calcium hydroxide, (c) trisodium phosphate, and (d) sodium carbonate.

Catalytic desulfurization is used to convert the sulfur to hydrogen sulfide, which is removed later by one of the above processes.

**Acid treatment.** This process removes the coloring materials from base stocks. Lubricating oils made from paraffin base crudes do not require acid treatment, while distillates from the mixed and asphalt types of crudes generally are refined with acid. A 93% solution of sulfuric acid (66 Beaumé acid) is most commonly used in acid treating. Sometimes a more dilute acid is used, especially when treating is done for color removal only; occasionally a 98% acid is used for lubricating stocks.

The amount of acid used will vary with the type of crude and with the specification of the product. For example: (1) natural gasoline, 2 lb of acid per barrel of product (the process is commonly run at 70–90°F); (2) straight-run gasoline, 3–5 lb of acid per barrel of product (70–90°F); kerosine, up to 20 lb of acid per barrel of product (90–130°F); lubricating oils, 0–50 lb of acid per barrel of product. The oils from Pennsylvania crudes require little or no acid treatment, mixed base crudes up to 20 lb/bbl and the asphaltic crudes require up to 50 lb/bbl (120–180°F).

The acid and the product being treated are agitated so that there is intimate contact between them. Some of the acid-treatment processes require 1 minute or less of contact time between the acid and the material being treated. Kerosines may require as much as 30 min contact time, while the lubricating stocks may require 1–2 hours. Continuous processes are in use today and contact time is being shortened considerably.

**Clay-contact adsorption treatment.** The use of clay treating for the purification of oils was known as early as 1822. Percolation filtration was ideally suited for the slight decolorization required by the Pennsylvania oils. The term percolation is applied to the filtering method in which the oil is passed through a bed of granular adsorbent clay.

Contact filtration makes use of a direct agitation of a very fine clay with the oil at elevated temperatures for a given time.

During the clay treatment the oils are neutralized and decolorized by the removal of the suspended matter. The decolorization of oil is an adsorption process in which the asphaltic and resinous chemicals of the oils are adsorbed on the surface of the clay particles.

The oil is mixed with the clay in the ratio of 20–80 lb of clay per barrel of oil (that is, from 5–30% of the weight of the oil). The commonly used clays are fuller's earth, bentonite, bauxite and alumina, and activated alumina.

After the oil and clay are mixed, the slurry formed is solvent-treated, using propane as solvent. After solvent treatment the resultant mixture is heated to the treating temperature (which may

**Table 5. Desalting and dehydrating methods**

Method	Temperature, °F	Type of treatment
Chemical separation	140–210	0.05–4% solution of soap in water 0.5–5% solution of soda ash in water
Electrical separation	150–200	10,000–20,000 volts
Gravity separation	180–200	Up to 40% water added
Centrifugal separation	180–200	Up to 20% water added (sometimes no water added)



range from 200–600°F depending on the nature of the original oil). The propane, moisture, and reaction products are removed in a vacuum stripper, and the oil is cooled and filtered. The oil is sent through additional filtering operations and then to storage. Before the oil is sold, additives are added to give the oils certain desired properties. The clay is reactivated in a clay regenerative furnace. See ADSORPTION.

**Vapor-phase treatment.** In this process, gum-forming compounds in the gasoline vapor are removed by adsorption onto clay materials. It is a low-cost treatment, as over 25,000 barrels can be treated per ton of clay before clay reactivation is necessary. The vapors from the cracking unit enter the treating tower (often called a polymerizer) near the top and flow downward through a packed clay bed. Polymerization occurs in the tower, and a polymer liquid containing a large percentage of gasoline collects at the bottom. The vapor passes to the condensers as finished gasoline except for the sweetening process. The polymer is drained from the tower and the gasoline recovered in a separate tower.

The use of gum inhibitors has to a great extent obviated this vapor-phase treatment.

**Solvent treatment in petroleum refining.** Undesired constituents may also be removed by selective solvent extraction. In this case a liquid that will selectively dissolve the undesired constituents is added to the oil. The solvent processes may be divided into two main categories, solvent extraction and solvent dewaxing.

The solvents used in the extraction processes include the following: propane and cresylic acids, 2,2'-dichlorodiethyl ether, phenol, furfural, sulfur dioxide, benzene, and nitrobenzene.

In the dewaxing process, the principal solvents are benzene, methyl ethyl ketone, propane, petroleum naphtha, ethylene dichloride, and sulfur dioxide.

Before the solvent-extraction processes were developed, only a few types of crudes were considered to be good "lubricating oil" crudes. By using these solvent processes the original properties of the crudes can be changed so greatly that almost any crude will make good lubricating oils.

The early developments of solvent processing were concerned with the lubricating oil end of the crude. Solvent-extraction processes are being applied to many useful separations in the purifications of gasoline, kerosines, diesel fuel, and others. In addition, solvent extraction may replace fractionation in many separation processes in the refinery. For example, propane deasphalting has replaced, to some extent, vacuum distillation as a means of removing asphalt from reduced crudes. See SOLVENT EXTRACTION; see also OIL ANALYSIS; PETROCHEMICAL; PETROLEUM; PETROLEUM PRODUCTS. [J.J.M.]

**Bibliography:** W. L. Nelson, *Petroleum Refinery Engineering*, 4th ed., 1958; *Petroleum Refiner Process Developments*, published September odd-

numbered years; *Petroleum Refiner Process Handbook*, published September even-numbered years; P. A. Washer, *Fundamentals of General Refinery Practices*.

## Petroleum products

Most crude petroleum is useful only as a raw material for the manufacture of a large number of products such as fuels, lubricants, paving material and base compounds for chemical manufacture. It is not unusual for a large oil company to list 200–1000 products for sale.

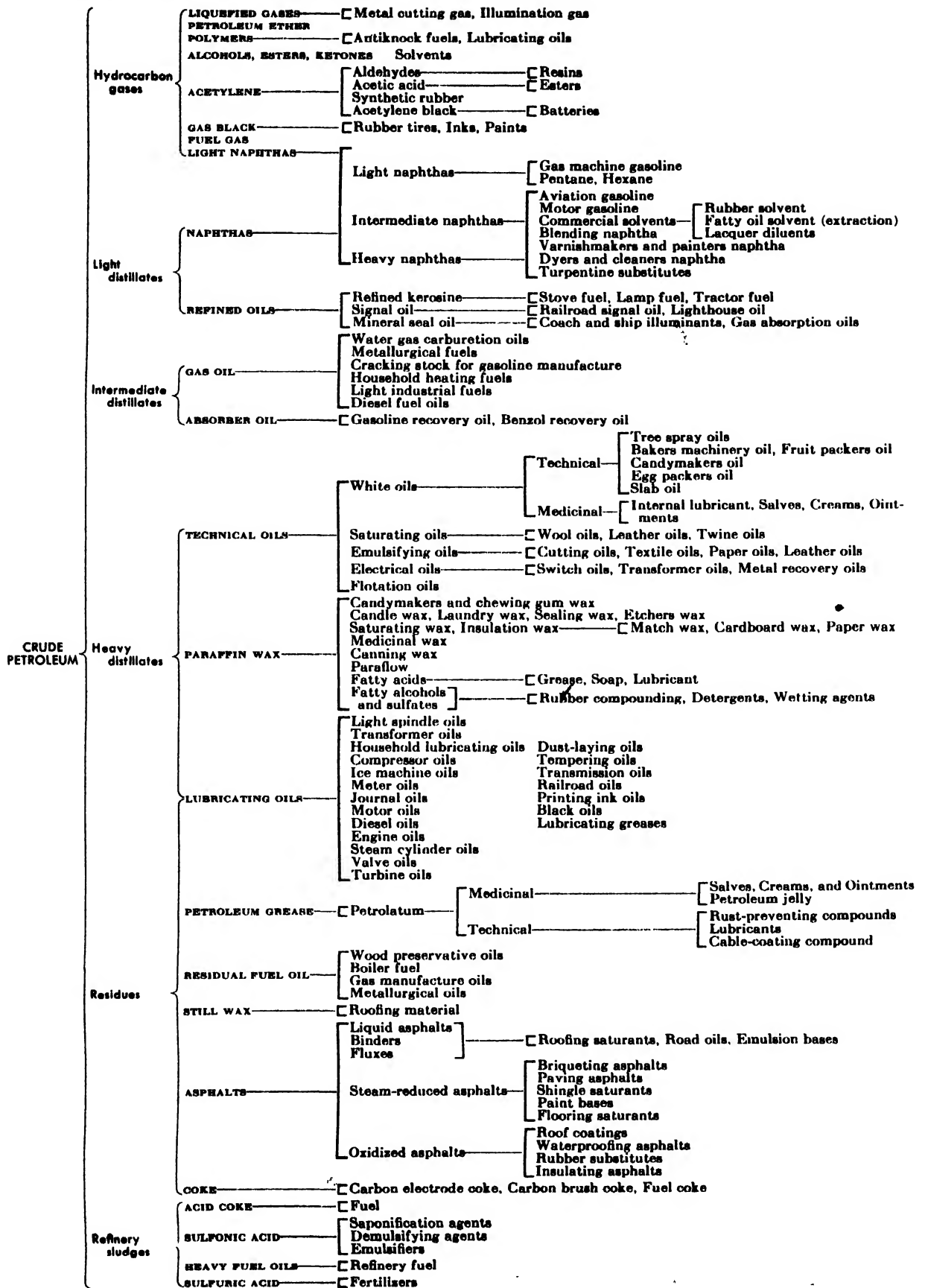
**Major products.** A normal crude oil will contain thousands of hydrocarbons, ranging in molecular weight from that of methane,  $\text{CH}_4$ , the chief constituent of natural gas with molecular weight of 12, to those of asphaltic compounds with apparent molecular weights of about 50,000–100,000. During refining, such a crude oil is first distilled into fractions which cover its boiling point range. The lowest-boiling fractions are the normally gaseous hydrocarbons, methane, ethane, propane, and the butanes. Methane is used as a fuel; the others are used as chemical raw materials or as constituents of liquefied petroleum gas sold in cylinders.

The hydrocarbons from pentane to dodecane ( $\text{C}_5$  to  $\text{C}_{12}$ ) go into gasoline motor fuel, and this product, in all its varieties, takes about 45% of the crude oil supply of the world. Gasolines for reciprocating engines cover a considerable range of chemical composition, with corresponding variations in properties, the more important being volatility and resistance to knocking. Minor uses of gasoline hydrocarbons are cigarette lighter fluids, spot removers, dry-cleaner's naphthas, and paint solvents.

Another 45% of the crude oil is used also for fuel manufacture: kerosine for lamps and stoves, diesel engines, and house-heating furnaces; mixtures of gasoline and kerosine for jet engines; gas oil for diesels, house heaters, and special industrial jobs; heavy residual oils for steam generation, for propelling ships, and for heavy industry such as cement and steel manufacture. The composition of the hydrocarbons in kerosine overlaps that of the heavy end of the gasoline fraction in hydrocarbon content, lying at about  $\text{C}_{10}$  to  $\text{C}_{14}$ ; the hydrocarbons in conventional gas oil extend from about  $\text{C}_{14}$  to  $\text{C}_{20}$ ; and heavy fuel oil may contain hydrocarbons up to  $\text{C}_{70}$  or  $\text{C}_{80}$ . Asphalts are composed of still more complex and essentially nonvolatile substances together with complex compounds of sulfur, nitrogen, and oxygen which are largely hydrocarbon in character.

**Minor products.** The above major products make up 90% of the output of the oil industry. The remaining 10% is divided into a wide variety of materials, the most important being the lubricants. These extend from very light, low-viscosity liquids employed for high-speed machinery such as spinning and weaving equipment, through very viscous oils for reciprocating steam engines to semisolid and even solid lubricating greases.

## Crude petroleum and some of its products\*



\* From P. Albert Washer, Texas A. and M. College Extension Division (First Session).

All petroleums except the condensate crudes (very light crudes carried as vapor in deep, high-pressure gas reservoirs) contain lubricating oil materials, but the lower-grade crudes, for example those of the Middle East, are contaminated with asphaltic materials and compounds of sulfur and nitrogen which cause difficulty and loss in refining. Good lubricating stocks can be made from such crudes with fair economy, using differential solvent extraction.

Before the development of solvent extraction methods, natural petroleums were favored for the production of lubricating oils, but they have undesirable characteristics which show up under severe operating conditions. Modern lubricants consist of a blend of refined petroleum products supplemented by additives—antioxidants, anticorrosion agents, dispersant-detergents, improvers of viscosity-temperature characteristics, and so on. Where liquid lubricants cannot serve, oils thickened with soaps of sodium, calcium, aluminum, and lithium or with oleophilic solids—modified silicas and bentonites—are employed.

Native asphalts are rarely satisfactory for use. They are often more expensive than those made from petroleum which have largely replaced the native materials because their properties can be manipulated by selection of suitable crude oils and by noncracking distillation, by oxidation at high temperature, by blending, and by use of additives. The quantities of asphalt employed for paving, roofing, and making molded articles are in the ratio 2:1:1.

Petroleum waxes, recovered in the making of lubricating oils from paraffinic crudes, fall in two classes: (1) refined waxes, macrocrystalline in type, essentially the normal paraffins  $C_{20}$  to  $C_{30}$ , used in candles, paper waxing, and household paraffin wax; and (2) microcrystalline waxes, the amorphous waxes of commerce, used widely in paper sizing, in coating frozen food packages, in insulation, and in making petrolatums.

Small quantities of petroleum are used to make technical white oils, emulsifier sulfonates, insulating oils, insecticides, rubber extenders, hydrogen and town gas, synthetic detergents, and intermediates for the chemical industry. The raw materials vary a good deal but are taken largely from the heavier fractions of the crude oil.

See separate articles on the more important products. *See also* ASPHALT AND ASPHALTITE; GASOLINE; LUBRICANT; PETROLEUM; PETROLEUM PROCESSING. [M.SO.]

## Petroleum reservoir engineering

The applied science concerned with the development and operation of reservoirs for maximum economic recovery of oil, gas, or both. It is a composite technology requiring coordinated application of many special scientific disciplines such as physics, geology, chemistry, and mathematics, as well as other engineering sciences, in the study of the complex reservoir systems.

The gross measures of a reservoir as an entity of commercial interest are (1) the amount of oil, gas, or both initially present in the reservoir; (2) the rates at which the hydrocarbons can be withdrawn; and (3) the fraction of the original hydrocarbons in place which can be economically recovered.

**Oil or gas in place.** The amount initially in place can be determined simply as the volume of reservoir rock containing the hydrocarbons of interest times the content per unit volume of rock. The former can be obtained by the thickness of the productive formation at wells drilled for its development, and the total productive area defined and outlined by these wells. The oil or gas content per unit volume of rock is essentially given by the measured or calculated porosity, reduced by the amount of water saturation; from the reservoir volume of the hydrocarbons the volume at the surface can be calculated. The two volumes differ because of shrinkage of crude oil on evolution of its solution gas or expansion of free reservoir gas as its pressure declines to atmospheric. The formation thickness, porosity, and connate or interstitial water saturation should be considered as locally variable, to the extent that such variations can be determined from information obtained at the individual wells.

If the total measured oil-bearing or productive area in acres is indicated by  $A$ ; the average formation thickness of the productive zone, excluding nonproductive members such as shales, in feet by  $h$ ; the average porosity by  $\phi$ ; the average water saturation by  $S_w$ ; and the formation volume factor of the oil by  $B_o$ ; the initial surface (stock tank) oil content in place  $N$  will be given by the equation

$$N = 7758.4Ah\phi(1 - S_w)/B_o \text{ barrels} \quad (1)$$

For the gas cap of an overlying oil reservoir, or a nonassociated gas reservoir, the gas content would be determined by the same equation, provided the term  $B_o$  is replaced by  $B_g$ , the volume at reservoir conditions of a unit volume of gas at the surface. The gas content in cubic feet is obtained by multiplying the calculated volume in barrels by the factor 5.6146. For a complete accounting of the gas content, that dissolved in the oil (the solution gas) is calculated as the oil content times the gas solubility at the initial conditions, expressed as cubic feet per barrel.

**The material balance equation.** The initial oil in place can also be inferred from observations on the pressure behavior within the reservoir as oil and gas are produced. Making a material balance for the gas and oil by interrelating the volumes produced with the amounts assumed to be initially present, it is found that  $N$  will be given by the material balance equation

$$N = \frac{N_p \left( \frac{R_p - R_s + B_o}{B_o} \right) - G_i - \left( \frac{1}{B_{gi}} - \frac{1}{B_g} \right) G - \frac{W_p}{B_w}}{R_{si} - R_s - (B_{oi} - B_o)/B_o} \quad (2)$$

where  $G$  = initial reservoir volume of free gas phase present

$W_e$  = net water intrusion volume

$N_p$  = cumulative oil production

$R_p$  = cumulative gas-oil ratio (total gas produced divided by  $N_p$ )

$R_s$  = gas solubility in oil

$G_i$  = cumulative gas injection, if any; subscript  $i$  elsewhere indicates initial values

In this equation,  $N$  and  $G$  are the basic unknown constants;  $N_p$ ,  $R_p$  and  $G_i$  are the actual quantities of production or injection; and  $R_s$ ,  $B_o$  and  $B_g$ , as well as their initial values, are functions of the pressure and can be determined by experiments with the oil and gas.  $W_e$  is, in principle, a variable unknown.

If it is known that the water intrusion term is not of an important magnitude and can be neglected, the material balance equation reduces to one with the two constant unknowns  $N$  and  $G$ . Application of the equation to two or more time periods for which the other terms are known will then permit its solution for  $N$  and  $G$ .

If  $G$  may be taken as zero, but the water intrusion term  $W_e$  is not an insignificant factor, calculations of  $N$  on ignoring  $W_e$  will show an increasing trend as production continues. This in itself will be strong evidence that water encroachment is playing a role in the pressure performance. Extrapolation of the calculated values of  $N$  to the time of initial production will often indicate reasonable values of the true magnitude of  $N$ . Conversely, if the latter is known or can be estimated independently the material balance equation can be inverted to calculate the volumes of water encroachment corresponding to the production performance.

**Darcy's law; permeability.** The rates at which the hydrocarbon fluids can be withdrawn from a reservoir depend on the number of wells draining the reservoir, the average thickness of the formation, and the inherent transmissibility of the reservoir rock for these fluids. The last factor is expressed by the term permeability. Its significance lies in that it is the coefficient of proportionality in the basic physical law governing the flow of fluids through porous materials, namely, Darcy's law (see FLUID-FLOW PRINCIPLES). In its generalized form, applicable to flow in a direction  $s$  inclined to the horizontal by the angle  $\theta$ , it may be expressed as

$$u = -\frac{k}{\mu} \left( \frac{\partial p}{\partial s} + \rho g \sin \theta \right) \quad (3)$$

where  $u$  is the volumetric rate of flow per unit area,  $\mu$  the fluid viscosity,  $\rho$  the fluid density,  $g$  the acceleration of gravity,  $\partial p/\partial s$  the pressure gradient, and  $k$  the permeability. If  $u$  is expressed in cc/sec,  $\mu$  in centipoises,  $\partial p/\partial s$  in atmospheres/cm and  $\rho g$  in atmospheres/cm, then  $k$  is in darcys.

The permeability unit, darcy, may be defined as the permeability of a porous medium which will carry a flow of 1 ml/(sec) (cm<sup>2</sup>) of a 1-centipoise

(cp) viscosity fluid under a pressure or hydraulic gradient of 1 atmosphere/cm.

In most practical applications it is convenient to express the actual permeability in the unit of the millidarcy, or thousandth of a darcy, (md). Consolidated producing sands generally have permeabilities in the range of a few to several hundred millidarcys. The permeabilities of unconsolidated sands and fractured or highly vugular limestones often range in the thousands of millidarcys. Tight productive limestones frequently have matrix permeabilities even lower than 1 md.

In the above differential forms of Darcy's law, the permeability  $k$  is to be considered as being variable from point to point in the medium, if the latter is not uniform throughout, even though the fluid itself persists as a single-phase liquid or gas. It may also have different values for different directions of flow. The primary variable is the pressure  $p$ . The validity of Darcy's law has been established by extensive experimentation, although, as most linear relationships do, it tends to break down if the fluid velocities are indefinitely increased. Within such limits, which encompass virtually all situations of practical importance, the flow is considered to be viscous.

The permeability defined above is independent of the nature of the fluid, provided it occupies the whole of the pore space, and depends only on the character of the porous medium. The viscosity and density alone suffice to discriminate between one fluid and another. At low pressures, however, permeabilities measured for gas flow have been found by L. J. Klinkenberg to be higher than those determined for liquid flow, but this effect is of minor importance except in laboratory experimentation on low-permeability materials.

Perhaps the simplest application of Darcy's law to a problem simulating one of actual oil and gas production relates to the steady-state horizontal flow into a well bore. Assuming the flow is radially symmetrical about the well, it is readily found that the pressure will increase as the logarithm of the distance from the center of the well. The rate of liquid flow is then given by the formula

$$q = \frac{2\pi kh(p_e - p_w)}{\mu B \ln r_e/r_w} \quad (4)$$

where  $k$  is the permeability,  $h$  the formation thickness,  $p_w$  the pressure at the well of radius  $r_w$ ,  $p_e$  the external pressure at radius  $r_e$ ,  $\mu$  the viscosity, and  $B$  the formation volume factor of the liquid.

In common practical units this equation becomes

$$q = \frac{0.003076kh(p_e - p_w)}{\mu B \log_{10} r_e/r_w} \text{ barrels/day} \quad (5)$$

where  $k$  is expressed in millidarcys,  $h$  in feet,  $\mu$  in centipoises, and  $p_e, p_w$  in psi. The external radius  $r_e$ , though not precisely defined, represents the area from which the liquid is being drained or that where the pressure may be assumed to be  $p_e$ .

For similar steady-state radial flow of gas into a well, the rate of flow  $q_g$  at 60°F and 1 atmosphere is found to be given by the equation

$$q_g = \frac{0.3056kh(p_e^2 - p_w^2)}{\mu Z T_r \log_{10} r_e / r_w} \text{ ft}^3/\text{day} \quad (6)$$

where  $Z$  represents the average supercompressibility or deviation factor of the gas in the reservoir as compared to an ideal gas, and  $T_r$  is the reservoir temperature in °R (Rankine).

**Multiphase production.** Actual systems always involve more than one fluid phase. Gas and oil and water and oil are the most common flow stream combinations. But even when gas and oil are flowing individually as single phases, the presence of the immobile connate water must be taken into account.

Experimental studies have shown that the multiphase flow through porous media can still be described by the basic Darcy type of equation, provided it is applied to each distinct phase separately and the associated permeabilities are considered to be functions of the fluid-phase saturations. Indicating the oil, gas, and water phases by the subscripts  $o$ ,  $g$ , and  $w$ , their simultaneous flow in the direction  $s$ , at angle  $\theta$  with the horizontal, will be governed by the equations

$$\begin{aligned} u_o &= -\frac{k_o}{\mu_o} \left( \frac{\partial p_o}{\partial s} + g\rho_o \sin \theta \right) \\ u_g &= -\frac{k_g}{\mu_g} \left( \frac{\partial p_g}{\partial s} + g\rho_g \sin \theta \right) \\ u_w &= -\frac{k_w}{\mu_w} \left( \frac{\partial p_w}{\partial s} + g\rho_w \sin \theta \right) \end{aligned} \quad (7)$$

The velocities  $u_o$ ,  $u_g$ , and  $u_w$  represent the volumetric flux rates of the corresponding phases. The pressures  $p$  are expressed individually, since they will change discontinuously across the curved interfaces between the phases. These pressure differences are referred to as capillary pressures, and may be considered as functions of the phase saturations, to be measured experimentally, although they are determined directly by the interfacial curvatures and interfacial tensions. They are often ignored in the treatment of large-scale systems, but they may be of importance near fluid fronts and in regions of rapid change of the fluid saturations.

**Effective and relative permeabilities.** The permeabilities  $k_o$ ,  $k_g$  and  $k_w$  are termed effective permeabilities. When expressed as fractions of the permeability for a single fluid phase, the absolute permeability, they are called relative permeabilities. The latter are always less than 1, reflecting the interference of each phase with the flow of the others. The saturations, of which they are functions, are expressed as the fractions of the pore space which they occupy, namely,  $S_o$ ,  $S_g$  and  $S_w$ , with a sum always equal to 1. The variation of the relative permeabilities with the phase saturations is referred to as the permeability-saturation relationship, and is illustrated by Fig. 1 for a mixture

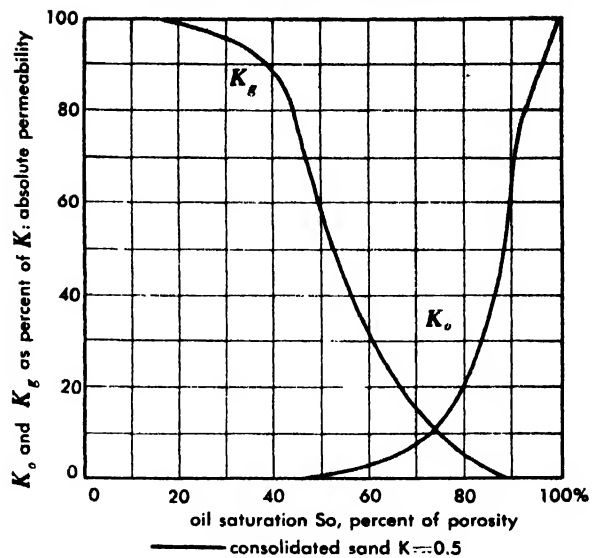


Fig. 1. Gas and oil relative permeability curves for a Nichols Buff sandstone. (After Botset, AIME Trans.)

of gas and oil flowing through a Nichols Buff sandstone.

In interpreting these curves it is helpful to distinguish between the gas as the nonwetting phase and the oil as the wetting phase, referring to their respective tendencies preferentially to adhere to and wet the internal solid surface of the rock. It then will be observed that the permeability for the wetting phase—oil in the case of Fig. 1—drops rapidly as its saturation decreases from 100%, and falls to zero long before its saturation vanishes. This drop is due to the fact that the initial desaturation of the wetting phase occurs in the larger pores, which contribute more to the permeability than their proportional volumetric content. Conversely, at higher degrees of desaturation of the wetting phase the latter is left in the finest pores and in disconnected flow channels—the irreducible saturation—permitting negligible flow capacity.

The nonwetting phase—gas—tends to remain in a discontinuously distributed state with zero permeability until sufficient saturation—the equilibrium saturation—is built up for continuity to be established. The larger pore channels so occupied then provide a rapidly rising permeability with increasing gas saturation. Virtually full single-phase permeability is achieved at less than complete liquid desaturation and while the smallest pores are still filled with liquid.

For the more general case where three phases—oil, gas and water—are flowing, it is found that whereas the permeability to the wetting phase—water usually—is determined only by its own saturation and qualitatively follows a curve such as that for oil in Fig. 1, the relative permeability to the oil or gas may depend on the distribution as well as amount of the other two phases. These effects are illustrated in Figs. 2 and 3, showing, in triangular plots, the results found for an unconsolidated sand. Curves of this type can indicate many of the gross

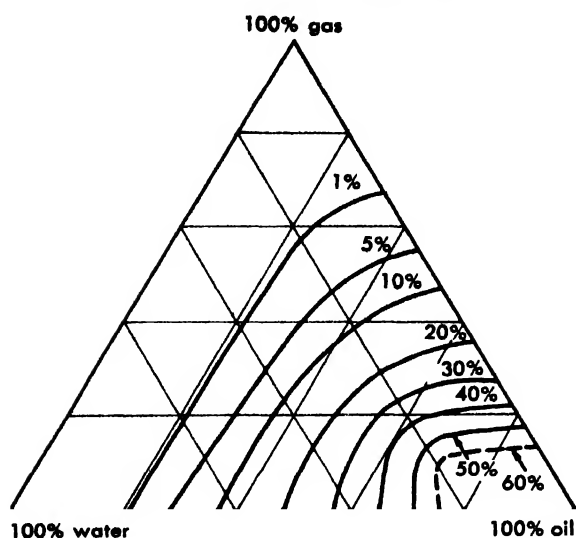


Fig. 2. Curves of constant oil relative permeability in flow of oil, gas and water through an unconsolidated sand as functions of the fluid saturations. (After Leverett and Lewis, AIME Trans.)

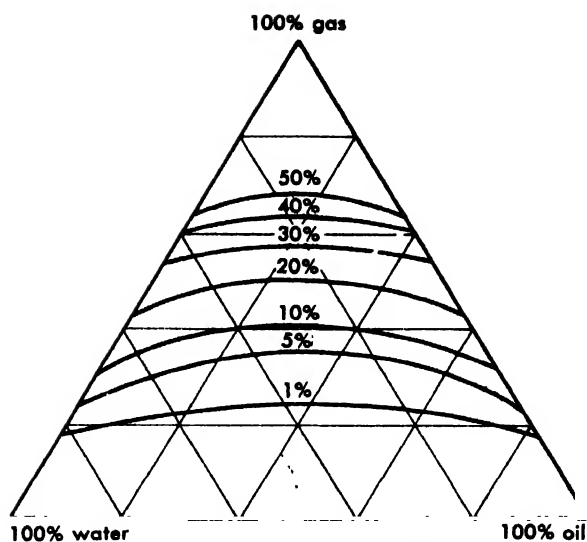


Fig. 3. Curves of constant gas relative permeability in flow of oil, gas and water through an unconsolidated sand as functions of the fluid saturations. (After Leverett and Lewis, AIME Trans.)

features of multiphase fluid flow in the porous medium of interest. For example, the nature of composite flow streams which can be maintained in different saturation ranges is illustrated in Fig. 4. It will be seen that simultaneous flow of all three phases in significant amounts will occur only in a very limited range of fluid saturations. By further reference to Figs. 1, 2, and 3 it will be observed that the composite permeability in multiphase flow will generally be but a nominal fraction of that for single-phase flow.

**Computing components of flow.** The actual fraction of any composite flow stream contributed by a particular phase can be calculated by combining the corresponding Darcy equations. For example, when gas and oil are flowing simultaneously the fraction  $f_g$  of the total volumetric flux  $q_t$  represented by the free gas phase is given by

$$f_g = \frac{\lambda_g}{\lambda_o + \lambda_g} \left[ 1 - \frac{\lambda_o}{q_t} \left\{ \frac{\partial P_c}{\partial s} - (\rho_o - \rho_g)g \sin \theta \right\} \right] \quad (8)$$

where the terms  $\lambda_o$ ,  $\lambda_g$  are the oil and gas phase mobilities, that is, the ratio of their permeabilities—the effective values—to their viscosities.  $P_c$  is the capillary pressure  $p_g - p_o$ . The corresponding fraction for the oil phase  $f_o$  is simply  $1 - f_g$ . For oil-water flow streams the oil and water fractions are given by the same equation after appropriate changes in the subscripts.

Relative and effective permeabilities are of importance not only in determining the detailed dynamics of the displacement of oil from reservoir rocks but also control the absolute flow rates. In the above equations for rates of production from wells, the permeabilities must be corrected for the connate water even if it is in its irreducible state and immobile, although the steady-state single-

phase flow equations will give only approximations of the actual flow magnitudes if both gas and oil are being produced.

**Energy and producing mechanisms.** Two basic, though elementary, observations underlie the essential principles of reservoir engineering. The first is that movement of viscous fluids such as oil through a reservoir rock involves the consumption of energy. Secondly, the withdrawal of oil from an oil reservoir requires a replacement of its volume in the reservoir space. Considered together these simple facts provide the framework for understanding the various types of oil-producing mechanisms.

Energy required for movement of oil from a reservoir rock into the producing wells may be drawn from four sources: (1) reservoir rock compression, (2) compression of reservoir and surrounding liquids, (3) compression of solution and free gas, and (4) gravity head above levels of withdrawals. Their individual importance depends not only upon the amount of such energy available but also on the effectiveness with which it can be used to displace the oil.

Upon release of the fluid pressure within the pores of a reservoir rock with removal of the oil and gas, the rock matrix will be subjected to increased compressive stress and compaction of the rock mass. However, in consolidated rocks the magnitude of such compaction within the reservoir itself will usually be too small to play a significant role in the oil expulsion processes. Observable and serious compaction has occasionally occurred in a few unconsolidated sand reservoirs, but in most cases the competence of the overburden apparently makes the compaction effects of minor importance.

The expansion of the connate water within a reservoir, assuming the water is undersaturated, will also generally be a minor factor in oil dis-

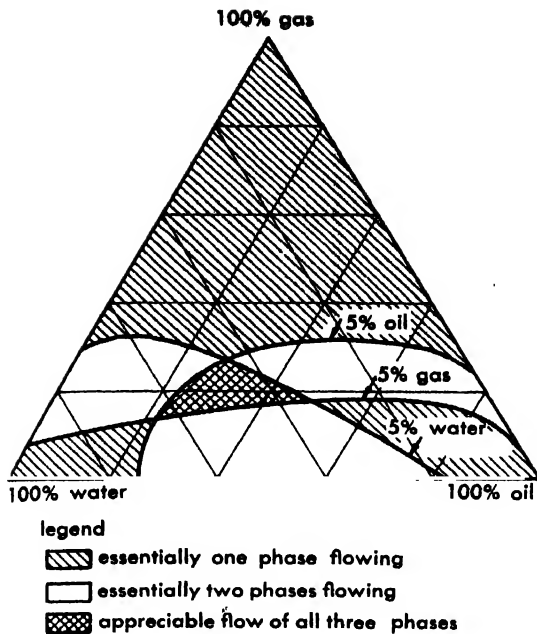


Fig. 4. Composition distribution of three-phase flow streams in an unconsolidated sand as functions of the fluid saturations. (After Leverett and Lewis, AIME Trans.)

placement. A pressure reduction even as high as 3000 psi will lead to an expansion of the water of only about 1%.

Reservoir crudes have compressibilities of the order of  $15 \times 10^{-6}$  per pound per sq. in. (psi), or about five times as great as water. Hence, if the reservoir oil is undersaturated by several thousand psi, its compression energy alone may provide an expansion in its volume and displacement into the producing wells of a few per cent of the total volume of the oil in place. Occasionally oil recoveries in this range may be of economic value. Usually, however, the expansion displacement of undersaturated oils is only a supplement to production controlled by other mechanisms.

**Water-drive reservoirs.** The compression energy of the water in aquifers which adjoin and are in communication with oil reservoirs is often the dominant energy source and mechanism for oil recovery. These are the water-drive reservoirs. Their important characteristics are the volumetric extent and the continuity of the aquifer and its ability to bring water into the oil reservoir fast enough to push the oil out at commercially profitable rates.

To the extent that the aquifers may ultimately outcrop at levels higher than the oil reservoir, the corresponding hydraulic head would, in principle, provide an artesian drive for displacing the oil. Because of the long flow path, however, the rates of flow through the aquifer as a whole will usually be too low to exert an appreciable influence on the reservoir production. On the other hand, it is the very large volumes of water in extended aquifers which make their compression energy and expansion potential important factors in oil displacement.

A circular 50-ft-thick aquifer of 50 miles equivalent radius will contain enough water to expand some 1,500,000 bbl for each psi of pressure reduction. Such levels of volumetric expansion could well support and replace the oil withdrawals from a great majority of oil-producing reservoirs. If the adjoining aquifer is of more limited extent or its large-scale continuity is interrupted by faulting or lithologic barriers, its contribution will be correspondingly limited.

A major phase of the study of a water-drive reservoir is the pressure and flow behavior of the aquifer itself. If the water compressibility and such contributory effects from the rock as may occur is taken as a constant  $c$ , the water density in the aquifer will be governed by an equation identical with the classical heat conduction equation, namely

$$\eta \nabla^2 p = \frac{\partial p}{\partial t} \quad \eta = k/\phi c \mu \quad (9)$$

To a very close approximation  $\rho$  can be replaced directly by the pressure  $p$ .

Solutions of the above equations analytically, by electrical network models, or by digital computers will show how the pressure in the aquifer will react to fluid withdrawals, which, in turn, can be related to the oil and gas production in the adjoining oil reservoir or the rates of water invasion into the latter. Conversely, from observed or assumed pressure histories at the water-oil boundary the rates of flow across the boundary and into the oil reservoir can be calculated. The geometrical and physical properties of the aquifer, about which advance information is often meager, can be determined in an empirical sense by trial and error adjustments so as to make the predicted pressure behavior match that observed in the course of producing the oil reservoir.

Many studies of this type and field observations show that the pressure in water-drive reservoirs is rate sensitive. That is, the average reservoir pressure depends not only on the cumulative oil production but also on the rates at which it has been withdrawn. Sharp increases in production rate will tend to accelerate the pressure decline per unit withdrawal. Cutbacks in withdrawal rate will generally retard the pressure decline and often even lead to buildups in reservoir pressure.

Water-drive reservoirs will permit maintenance of high rates of total fluid withdrawals through most of the economic life, although increasing volumes of water production will continually reduce the net oil rates. The reservoir pressure will tend to stabilize after initial declines which are necessary to induce the water to flow into the oil zone at sufficient rates to replace the oil withdrawals. Gas-oil ratios will rise only moderately during the producing life and in relation to the decline in reservoir pressure.

The water-drive mechanism is of special importance because of the high oil recoveries it often yields. Recovery factors as high as 50% of the initial oil in place are not uncommon, and under very favorable conditions they may be as high as



80%. The main factors controlling the recovery are the uniformity of the oil reservoir body and the viscosity of the reservoir crude. Variations of the permeability in the producing formation may lead to channeling of the invading water through high permeability zones and premature drowning out of the producing wells, so that while the oil displacement efficiency may be high in the invaded strata, the over-all average sweep efficiency and recovery will be relatively low. The viscosity of the reservoir crude controls the local displacement efficiency. The latter will be reduced as the oil viscosity increases. Because of this factor water-drive recoveries in reservoirs producing oils of gravity lower than about 20 degrees API may be considerably less than the 50% frequently observed for high gravity producing reservoirs.

The water-drive producing mechanism controls the production in important reservoirs in all major oil provinces. Many of the large reservoirs in Texas along the Gulf Coast operate under water-drives, as does the East Texas Field, the largest in the country. The two main sands in the gigantic Burgan Field in Kuwait are virtually perfect water-drive reservoirs.

**Solution-gas drives.** The gas dissolved in reservoir crudes is the most common energy source and displacement medium involved in oil production. When it is the dominant agent for oil recovery the producing mechanism is termed the solution-gas drive or depletion drive. The decline in reservoir pressure, which necessarily follows any appreciable production of oil and gas, will lead to liberation of solution gas within the pores of the rock and corresponding replacement of the volume of reservoir fluid withdrawn, if the oil is gas-saturated at the initial pressure, as it usually is. If the adjoining aquifer does not then supply an influx of water to provide for continued replacement of the oil withdrawals, the pressure will keep on falling, with continued additional evolution of dissolved gas. Ultimately the pressure and the dissolved gas will be dissipated and the economic life of the reservoir will be terminated.

The reservoir pressures and displacement processes in dissolved gas-drive systems basically are not rate sensitive. They depend only on the reservoir volume of total fluid withdrawals, although the rate and manner of oil production may affect the relative amounts of gas and oil produced and hence the total composite voidage for fixed quantities of oil recovery. As the evolved gas builds up the gas saturation, the permeability to the gas will grow, facilitating its escape to the producing wells without a corresponding increased displacement of the reservoir oil. As a result, after an initial period of rather constant gas-oil ratios, at the level of the initial solution value, the ratio will rise steadily to peaks of the order of 10-20 times as great, and then decline as the contribution of the free gas falls with decreasing pressure. The well- and field-producing capacities will also fall because the driving reservoir pressure drops and the permeability to the oil is reduced with increasing gas saturation.

Except for local effects about the producing wells themselves, the over-all history of depletion of a gas-drive reservoir can be predicted by the equation

$$\frac{dS_o}{dp} = \frac{\alpha S_o + (1 - S_o - S_w)\epsilon + \zeta S_o(\kappa - rR/\gamma)}{1 + \frac{\mu_o}{\mu_g}(\kappa - rR/\gamma)} \quad (10)$$

$$\text{where } \alpha = \frac{B_g}{B_o} \frac{dR_g}{dp}; \quad \epsilon = \frac{-1}{B_o^2} \frac{dB_o}{dp}; \quad \zeta = \frac{\mu_o}{\mu_g B_o} \frac{dB_o}{dp}$$

$$\text{and } \gamma = \frac{\mu_o B_o}{\mu_g B_g}; \quad \kappa = \frac{k_g}{k_o}$$

$r$  is the fraction, if any, of the produced gas which is returned to the reservoir.  $R$ , the current gas-oil ratio, can be related to the other variables as

$$R = R_s + \gamma \kappa \quad (11)$$

In these equations  $\alpha$ ,  $\epsilon$ ,  $\zeta$ ,  $\gamma$  and  $\mu$  are all functions of the pressure  $p$  determined by the properties of the gas and oil. The pertinent rock characteristics enter through  $\kappa$ , the ratio of the gas to oil permeability, as expressed as a function of  $S_o$ . The solution of Eq. (10) will show how the current oil saturation  $S_o$ , in the reservoir declines with falling pressure. It will also give the gas-oil ratio  $R$  at the corresponding period. The associated total oil recoveries per acre-foot of productive reservoir at any stage of depletion will be found by

$$N_p = 7758.4 \phi \left( \frac{S_{oi}}{B_{oi}} - \frac{S_o}{B_o} \right) \quad (12)$$

Solutions of Eq. (10) give the typical performance relationship of reservoir pressure and gas-oil ratio versus cumulative production as is observed in actual producing fields. By its construction, in which local well bore effects are ignored, it does not provide for any rate sensitivity of the recoveries.

Except for the mechanism of undersaturated reservoir oil expansion, solution-gas drives are the most inefficient producing systems. This is not because of the lack of sufficient solution-gas energy, but rather because of the internal bypassing of the gas as its saturation is built up so as to escape from the reservoir at high gas-oil ratios and little displacement effectiveness. The increasing oil viscosity, as the pressure declines and the solution gas is evolved, aggravates this effect. Ultimate economic recoveries are 10-30% of the initial oil in place, decreasing generally as the API crude gravity decreases or as the oil viscosity increases.

Solution-gas-drive recovery has been the dominant producing mechanism in many of the older fields developed in the United States in the mid-continent area, West Texas, and California. In recent years appreciation of the low recovery efficiency of this drive has led to the application of fluid injection operations or limitation of the rates of production so as to facilitate potential water drives or gravity segregation assuming greater roles in the recovery mechanism.

**Gravity-drainage drives.** Gas caps overlying an oil zone contain additional compression energy for oil displacement to supplement that of solution gas (see PETROLEUM GEOLOGY). If, as may happen under high production rates and pressure differentials, this gas is permitted to break into the oil zone and join the solution gas flow stream it will be dissipated rapidly and will result in rather limited increased oil recoveries. Its effect will be similar to that of dispersed gas injection directly into the oil zone. If, however, the pressure gradients in the reservoir are so restricted as not to overbalance the gravity differential between the gas and oil, the gas cap will be preserved as a segregated driving piston on the oil zone.

Simple downward drainage by gravity of oil in a vertical column of a porous medium will lead to low residual oil saturations and high recoveries, limited only by the permeability, wettability, and capillary pressure characteristics of the rock. In actual reservoirs, with or without gas caps, it is generally not feasible to simulate pure gravity drainage because the corresponding rates of production will be too low for maximum economic return. The inherent downward flow capacity of the rock will be further restricted by the decreasing permeability to oil as the pressure declines and the solution gas is evolved.

In practice, when the upper part of a reservoir trap contains a gas cap the compression energy of the gas is permitted to supplement that of the gravity head so as to provide rates of withdrawal at economic levels. The gas cap also serves as a surge chamber to retard the pressure decline and hence lessen the rate of gas evolution within the oil zone and the associated effects of reduced oil permeability and increased oil viscosity.

The displacement effectiveness of the expanding gas on the underlying oil zone is decidedly rate sensitive, as may be inferred from Eq. (8). The gravity-drainage mechanism as a whole is likewise affected by the rates of the displacement processes, and becomes less efficient as the latter increase. To achieve the high recovery potential of the gravity-drainage mechanism in a gas cap reservoir, a balance must be made between the beneficial use of the driving pressure in the gas cap to support the desired levels of downflank production and the simultaneous deterioration in the displacement efficiency where the gas has invaded the oil zone. When such a balance is achieved, the gas cap will appear to expand downward as a piston with a relatively sharp gas-oil contact transition zone.

Under favorable conditions of gravity-drainage operations, the upward buoyancy force on the gas evolved within the oil zone will overcome the downflank pressure gradients and the gas will migrate upstructure into the gas cap while the oil is flowing downward. Such countercurrent gas migration will aid in maintaining the reservoir pressure as a whole as well as high levels of oil saturation and permeability in the oil zone. Even when there is no initial gas cap this process can lead to the formation of secondary gas caps with subsequent be-

havior essentially similar to that of a primary gas cap.

The pressure in gravity-drainage drives, in which effective segregation between the gas and oil is achieved, will decline slowly. Production rates and capacity will hold rather steady except that upstructure wells will be successively shut in as their producing levels are reached by the expanding gas cap. The gas-oil ratios will follow the trend of the solution ratio if downward gas coning is not permitted and the evolved solution gas is allowed to migrate into the gas cap.

To promote the general benefits of maintenance of pressure and production capacity part or all of the gas produced in gravity-drainage reservoirs is often returned to the reservoir through injection wells completed in the gas cap. If enough gas is injected to replace the reservoir withdrawals fully and prevent any pressure decline, the maximum potential of gravity drainage can be achieved, provided it is not nullified by excessive production rates and gas breakthrough. In any case, the higher pressure levels at which the reservoir is depleted will mean that whatever residual reservoir oil does remain undisplaced will have higher shrinkage and will represent less unrecovered stock tank oil than if the pressure had not been maintained.

It is preferable that the gravity-drainage mechanism, where potentially available, be allowed to function throughout a reservoir's producing life. But even when this is not feasible, gravity drainage may still serve to prolong the economic life by resaturating the lower part of the oil zone after its rapid depletion by solution-gas drive. The long persisting settled production of old fields which have lost their pressure and reservoir gas often reflects the emergence of gravity drainage as a residual source of energy for bringing the oil into the well bores.

The main requirements for the effective development of the gravity-drainage mechanism are high structural relief, long oil column, and good vertical permeability or mobility. When these are present and full advantage is taken of them, recoveries as high as 70-80% of the original oil content can be achieved. Proportionately lower recoveries will be obtained when gravity drainage merely supplements the solution-gas-drive mechanism.

Gravity drainage, often aided by gas injection at the structural crest, has played an especially important role in the production of many of the oil reservoirs in Eastern Venezuela. A number of large fields in West Texas and in the Gulf Coast have benefited from gravity drainage. Several of the major fields in Iran also appear to operate with significant gravity segregation.

**Reservoir engineering analysis.** The primary starting point for the analysis or prediction of the performance of an oil reservoir is its geological structure and environment. This information can only be satisfactorily obtained from wells drilled within the areal confines of the reservoir or its immediate vicinity. Geological, electrical, and radio-

active logs and the study of cores of the productive rock itself provide the basic data. These, plus determination of the properties of the oil and gas, may suffice to determine the total initial oil and gas contents of the reservoir by applying volumetric Eq. (1).

The real reservoir secrets unfold after the reservoir is placed on production and observations are made on its performance—the history of its production of oil, gas and water, of its pressure, and the distribution of these among the various producing wells. These data combined through the material-balance Eq. (2) may give further checks on the initial fluid contents as well as indications of the relative roles being played by the various producing mechanisms.

Quite often at least two or all major types of producing mechanism will contribute appreciably to the composite reservoir behavior, and its analysis will require setting up equations for combination drives. Partial water drives actually occur more frequently than complete water drives. As previously indicated, gravity drainage usually supplements the solution-gas-drive processes, at least to some extent. Even in water drives gravity segregation may be of benefit in minimizing channeling and water coning effects and thus improving the overall sweep efficiency.

The ultimate recoveries are determined by the magnitude of the average residual oil saturation when production is terminated.

**Recent developments.** Injection of gas or water to supplement the native energy and oil displacement potential of the reservoir in its original state have become established practices. Such fluid injection operations are now generally undertaken early in the producing life of the reservoir, and as soon as it is determined that otherwise the recoveries will be limited to the inefficient levels of solution-gas drive or reservoir liquid expansion. For the older fields which were substantially depleted before the desirability of pressure maintenance was appreciated, secondary recovery installations have often been made in the form of gas repressuring or water flooding.

The ultimate limitation of oil recovery by presently established methods lies in the fact that the fluids—gas or water—which serve as the displacing phase are basically immiscible with the oil. Their surfaces of contact are therefore well defined interfaces. Because of the tremendous interfacial area thus distributed throughout the microscopic pores of the reservoir rock, these represent correspondingly large total capillary forces and energies. Except for their beneficial action in inducing imbibition of water in water-wet systems these capillary forces offer resistance to multiphase flow at all saturations, and tend to break up any flowing phase into a discontinuous and immobile distribution as soon as its saturation falls to critical limits. When the latter state is reached, the capillary forces hold the residual oil unrecoverable in spite of continued passage of immiscible displacing

fluids such as gas or water. These interactions are empirically expressed by the relative permeability-saturation relationships.

If the oil were displaced by a miscible fluid, the interfacial and capillary forces would be eliminated and local displacement efficiencies approaching 100% would result. This principle has long been applied in cycling gas-condensate reservoirs. Here dry liquid-stripped gas is injected into the formation to displace the condensate-containing reservoir gas and at the same time prevent declines in pressure and retrograde condensation and loss of its liquid content. Both gases are mutually miscible and the displacement proceeds without interface formation and capillary forces.

In the case of an oil reservoir, displacement by a miscible liquid such as the liquefied petroleum gases—propane and butane—would achieve similar results. However, to circumvent the economic burden of refilling the whole oil reservoir with these salable liquid products only a relatively small buffer zone or slug of the latter is used—up to 10% of hydrocarbon pore volume—and it in turn is displaced by gas. At pressures of the order of 1500 psi or greater, natural gas will also be miscible with the intermediate hydrocarbons or liquefied petroleum gases, at reservoir temperatures. Thus a continuous phase transition is developed, without the interfaces and capillary forces, from the reservoir oil to the miscible slug and to the final gas displacement phase. Field tests of this method of oil displacement and related modifications are now underway.

A quite different type of technique for improving oil recovery is that of in-situ combustion. Though suggested many years ago, it has been studied and developed on the basis of modern reservoir engineering principles only recently. It appears to have special promise in application to heavy oil reservoirs where because of high reservoir oil viscosity and very unfavorable mobility ratios for displacement by gas or water the latter conventional recovery methods are of low and often noncommercial efficiency.

In essence in-situ combustion, as investigated and tested to date, consists of the injection of air into the producing formation to sustain burning of the oil in place—at temperatures of the order of 600°F—and provide a flow of heat ahead of the combustion zone to lower the oil viscosity and increase the recovery and producing well productivity. The burning of the oil generates combustion product and vaporized oil gases. These, together with the bank of condensed water vapors, form a composite gas and water drive, moving toward the producing wells with the combustion front. In addition, the gases carry heat and raise the temperature of the rocks and fluids ahead of the combustion front, although the rapid attenuation of the temperature wave tends to delay the improvement in well productivity until the fire comes close to the producing wells. The vaporization process immediately ahead of the burning front leaves de-

posits of heavy oil residual or coke, and these serve as fuel for the final combustion reaction. As a result the rock through which the fire passes is left essentially clean with all its oil displaced or burned out. About 15% of the original oil may be consumed in this manner, some 85% thus being in principle recoverable in the rock traversed by the fire. Because the heaviest components of the oil are used as fuel, there is a tendency for improvement in the gravity of the oil recovered by in-situ combustion.

The fire may be started by heaters or heating processes developed in injection wells or by spontaneous combustion of the reservoir crude resulting from the exothermic oxidation and absorption of the oxygen in the air stream. The operations are carried on in pattern distributions of injection and producing wells similar to those used in water flooding.

Commercial success of in-situ combustion requires relatively high porosity and oil saturation so as to hold the ratio of air injection to oil produced down to economic levels. It is also necessary to develop enough gas permeability through the formation to permit sufficient through-flow of air and combustion gases to sustain the burning. To minimize the effect of heat losses to the top and bottom bounding strata, the reservoir bed must be of appreciable thickness. As in all displacement processes, uniformity of the producing section will facilitate achieving high sweep efficiency. However, the rapid advance of the burning front through a limited zone, by gravity segregation of the air and gases or by permeability channeling, will accelerate the transmission of the direct thermal effects to the oil masses near the producing wells, which would otherwise retain their high viscosity and low mobility until late in the recovery life of the reservoir.

A number of field trials have confirmed the basic feasibility of carrying on in-situ combustion in oil reservoirs. At the South Belridge Field in Kern County, California, some 50% or more of the 13-degree API oil in place in the test area was recovered within 18 months after air injection was started.

The various improvements in oil recovery processes discussed here represent only the major developments which already appear to have some range of economic feasibility. It is to be expected that not only will these and the older methods be materially improved by continued research, but that still more novel and powerful techniques will be brought to light as the science of reservoir engineering is perfected.

See OIL AND GAS FIELD EXPLOITATION; OIL AND GAS STORAGE; OIL AND GAS WELLS; PETROLEUM SECONDARY RECOVERY. [M.M.]

**Bibliography:** M. Muskat, *The Flow of Homogeneous Fluids through Porous Media*, 1946; M. Muskat, *Physical Principles of Oil Production*, 1949; S. J. Pirson, *Oil Reservoir Engineering*, 2d ed., 1958.

## Petroleum secondary recovery

The process of removing oil from its native reservoir by the use of supplemental energies after the natural energies causing oil production have been depleted. Petroleum secondary recovery contrasts with primary recovery, which is the oil production resulting from indigenous reservoir energies. Advancing principles of technology and conservation demand that natural energy be supplemented soon after discovery of a reservoir; therefore today's best practices combine the primary and secondary recovery periods. However, there are many reservoirs which have been depleted without benefit of supplemental energy, and in the narrowest sense, secondary recovery applies to the further development of these depleted reservoirs.

Secondary recovery was first practiced in the older and shallower petroleum reservoirs of the Appalachian region. It has since spread to all oil-producing regions of the world and has doubled the producing life of some oil fields. When combined with primary recovery, it is applied to deep reservoirs. When practiced on depleted reservoirs, it is generally limited by economic factors to reservoirs shallower than 3000 ft. Total oil production in the United States by secondary recovery methods in 1954 as compiled by the Interstate Oil Compact Commission was approximately 480,000,000 bbl.

**Energy supplement and well patterns.** Energy is supplemented by introduction of either gas or water under pressure into the reservoir. The use of gas is commonly known as gas-drive or gas-repressuring, the use of water as water-flooding. The injected fluid drives the oil remaining in the reservoir to the vicinity of production wells, from whence it can be lifted to the surface, and also takes up the space within the reservoir previously occupied by the oil.

Two types of wells, injection wells and production wells, are required for secondary recovery operations. Standard patterns have evolved for the arrangement of these wells. Locating wells in patterns permits intensive development of a given land area and ensures the maximum penetration of injected fluid to all parts of the reservoir. In the early history of secondary recovery, a single injection well was surrounded by a large number of production wells. This pattern, known as a circle drive, is still used for gas injection operations. Another pattern, the line drive, is a line of injection wells offset by a line of production wells.

The most common well pattern is the five-spot. Square networks of injection wells and production wells interlock so that each injection well is at the center of a square consisting of four production wells, and each production well is at the center of a square consisting of four injection wells. Another standard pattern is the seven-spot, which consists of injection wells located at the vertices of hexagons with a production well in the center of each hexagon. By exchanging the roles of injection and production wells in the seven-spot pattern, the

four-spot pattern is obtained. Nine-spot patterns are five-spot patterns with additional injection wells added at the midpoints of the sides of each injection well square.

The spacing between injection wells and production wells will depend upon local physical conditions of the petroleum reservoir and upon economic factors. The resulting well densities will range generally from one well per acre to one well per 40 acres. Spacing economics is controlled by the amount of gas or water that can be injected into or produced from a single well. The prediction of the amount of fluid which an injection well will handle is therefore one of the most critical technical points in planning a secondary recovery project. This amount of fluid will depend upon factors such as the permeability of the reservoir rock, the viscosity of the reservoir oil, the fraction of the reservoir pore space that is filled with oil, the thickness of the reservoir formation, the reservoir pressure, and the available surface pressure.

**Factors of effectivity.** The efficiency of a secondary recovery operation is determined by the effectiveness with which the injected fluid displaces oil from that part of the reservoir which it invades, and the degree to which the injected fluid can be made to invade all parts of the reservoir.

*Displacement and retention factors.* Even under the most favorable conditions of fluid invasion, it is not possible to replace all the oil in a given segment of reservoir rock. The rock contains a complex and interconnecting assemblage of small channels which are not uniform in shape or in size. Hence it is possible for the invading fluid to bypass and trap some of the oil-containing channels or oil globules. This residual oil is held in place by the strong capillary forces that are operative. On the basis of its oil-retentive properties, a reservoir rock may be classed as either oil-wet or water-wet. In the former, the residual oil may be held as a film or as filling the most minute pore spaces. In the latter, the residual oil may be held as trapped globules or islands within the larger pore spaces.

*Fluid segregation problems.* Because oil is less dense than water and more dense than gas, there may be a segregation of injected water to the bottom part of the reservoir, or of injected gas to the top part. In either case, the injected fluid will advance toward the production well through only part of the reservoir. Complete entry of the invading fluid to all parts, then, is not possible without production of large amounts of injected fluid from the production well, a procedure which is necessarily costly. The possibility of encountering fluid segregation may lead to the deliberate locating of wells so that injection of gas is to the top of a reservoir structure or injection of water to the bottom, thus making use of the segregation tendencies. This cannot be done, however, in flat, thin reservoirs.

Reservoir rock properties are seldom uniform, and in particular, the rock may vary in its perme-

ability, that is, in its capacity to conduct fluid. The injected fluid will take the path of least resistance, and by the time it has invaded the higher permeability channels completely, it will have invaded the lower permeability channels only partially. If the permeability channels are sufficiently stratified, special well-completion techniques can be used to promote uniform fluid invasion.

If the resistance to flow is higher for the oil in the reservoir than for the invading fluid, the invading fluid will seek the production well by the most direct flow path. As a result, injected fluid will reach the production well before all of the pattern area has been invaded. The area of a pattern which has been invaded by the time the invading fluid breaks into the production well is termed the areal coverage of the pattern.

*Mobility ratio and area extent.* The ratio of flow resistances between injected fluid and reservoir oil is known as the mobility ratio. With a mobility ratio of unity, the area coverage for the five-spot pattern is 72.3%; for the seven-spot and four-spot, it is 74%. Each pattern has a characteristic coverage for each mobility ratio. When the flow resistance of the reservoir oil is extremely low, the coverage will approach unity for any type of pattern.

As the flow resistance of the reservoir oil increases in comparison to that of the invading fluid, the achieved areal coverage will decrease. For this reason, injected gas generally produces a smaller areal coverage than injected water. However, because of the more favorable economic factors for handling gas, the injection of gas can be continued for long periods of time after gas arrives in the production wells. After water arrives at the production well, water injection generally can be continued only until the ratio of water-oil volume reaches 1:100 or less.

*Oil recovery and residue.* The percentage of oil recovered by secondary methods will be a composite result of the effects that have been noted, namely, of the geometrical arrangement of pores and the capillary forces, of the gravity segregation, of the heterogeneous nature of the reservoir, and of the areal coverage that can be achieved. Under the most favorable circumstances, one may expect residual oil following secondary recovery to be as low as 15–20% of the pore space. With some of these factors operating to a disadvantage, residual oil following secondary recovery may be as high as 50% of the pore space. The residual oil will be highest where rocks have complex porous structure, where the reservoir oil is quite viscous, or where there are wide variations in reservoir permeability.

**Hazards from fluid contamination.** In water flooding, considerable attention must be given to the purity of injected water. The reservoir rock will act as a filter to remove suspended material in the well bore and clog the formation. Hence, no suspended material can be permitted. Even bacteria will filter on some rock formations and reduce the

injection rate. Consequently, treatment for removal of bacteria will often be necessary. Highly corrosive waters must also be avoided.

There is always the possibility that chemical interactions will occur between ions contained in the injected water and those present in the native reservoir waters or reservoir minerals. Injected water may also produce a change in the structure of reservoir clay material with a resultant reduction of flow capacity.

In the injection of gas, principal attention is paid to the removal of materials that might condense within the reservoir, produce corrosive action on operating equipment or yield oxidation within the reservoir.

**Experimental developments.** Advancing technology has continued to seek means of improving secondary recovery. Principal among these efforts has been the attempt to remove completely the capillary forces that are operative in retaining residual oil. This can be accomplished in the laboratory by a process known as miscible displacement, which consists of replacing the reservoir oil by a fluid with which it will mix completely. For example, one might inject a solvent, such as butane, until reservoir oil had been completely dissolved and brought to the surface. Because it is not economical to use the large amounts of solvents required for this process, techniques have been suggested for using slugs of solvent followed by water or gas, the character of the solvent being such that it will mix freely with the following water or gas as well as with the oil which it is to replace. Alcohols such as tertiarybutyl alcohol have been proposed.

With gas injection, a miscible slug zone can be achieved between the oil and following gas by using injection pressures in the range of 3000–4000 psi or by adding certain gas components to the injection stream.

Another method for reducing capillary forces is the addition of surface-active materials to a water flood, removing oil much as one would remove grease with a detergent. The principal disadvantage of using surfactants is that they are adsorbed on the reservoir rock. Their use is not economic unless means can be found for continually desorbing or replacing the surfactant. Other proposed additives in water flooding are soluble gases such as carbon dioxide.

Other experimental methods are the thermal and in-situ combustion methods. A fire, or combustion process, is started in the reservoir at an injection well. By the continued introduction of gas containing oxygen or other material to support combustion, the combustion wave is driven through the reservoir toward the production well. As the combustion wave moves forward, part of the oil is distilled and driven forward; part of the oil is burned to produce the heat necessary for continuing the combustion drive.

Much laboratory research has been done on these proposed newer methods of secondary re-

covery, and numerous field tests are now under way to examine all of these methods. None of them can be said to have reached the stage of technology where their effective use can be predicted, or where the factors controlling the economics can be delineated.

See OIL AND GAS FIELD EXPLOITATION; PETROLEUM RESERVOIR ENGINEERING. [J.C.C.]

**Bibliography:** American Petroleum Institute, *Secondary Recovery of Oil in the United States*, 1950; J. C. Calhoun, Jr., *Fundamentals of Reservoir Engineering*, 1953; M. Muskat, *Physical Principles of Oil Production*, 1949; L. C. Uren, *Petroleum Production Engineering: Oil Field Development*, 4th ed., 1956.

## Petrology

The study of rocks, their occurrence, composition, and origin. Petrography is concerned primarily with the detailed description and classification of rocks, whereas petrology deals primarily with rock formation, or petrogenesis. A petrological description includes definition of the unit in which the rock occurs, its attitude and structure, its mineralogy and chemical composition, and conclusions regarding its origin. In a restricted sense, however, petrology has come to emphasize the study of rocks in the field and in hand specimens, without recourse to the microscope. For a discussion of mineral identification, petrographic analysis, and the classification of rocks see MINERALOGY; PETROGRAPHY; ROCK.

**Igneous rocks.** Extrusive (effusive) igneous rocks reach the surface either through fissures of considerable linear extent (fissure eruptions) or through pipelike channelways around which volcanoes are built. Extrusive material may flow out relatively quietly as lava or it may be exploded as pyroclastic material. Fissure eruptions are generally quiet and repeated over long periods of time to build up thick platforms of considerable extent consisting chiefly of basalt. In the northwestern United States the Columbia River Plateau, built in this way, embraces 200,000 mi<sup>2</sup> in Idaho, Oregon, and Washington, and in local areas has an aggregate thickness of 5000 ft. See VOLCANO.

Volcanic structures are of a variety of types (see Table 1). Lava flows may be characterized by a smooth or ropy surface with prominent flow structure (pahoehoe), or by a jumbled blocky surface (aa). Flows commonly show columnar jointing which has been produced by contraction upon crystallization. A lava tongue solidifies first along its upper surface against the air and along its bottom contact with cooler rock, leaving a central stream which is still liquid, flowing in a tunnel of its own construction. With sufficient slope the streams drain away, leaving cavernous passageways.

Volcanic activity varies greatly in intensity, duration, periods between eruptions, and quantities of gases, liquid rock, and solidified fragments expelled. The more important factors influencing



these differences are (1) chemical composition of the magma; (2) amount of gas dissolved in it; (3) extent of crystallization or cooling before eruption; and (4) configuration of the conduit and depth to the magma chamber. *See* MAGMA.

Intrusive igneous rocks occur in many different types of units or intrusive masses, which are classified chiefly by their shape and structural relations to their wall rocks (Table 2). Bodies that crystallized at great depths (such as batholiths) are referred to as plutonic; those consolidated under shallow cover are designated as hypabyssal. *See* PLUTON.

The crystallization of the larger intrusives may result in profound alterations in the adjacent wall rocks (exomorphism). Where stocks and batholiths have invaded sedimentary rocks an aureole of contact metamorphism is developed. This results from recrystallization under increased temperature and may be accompanied by chemical transformations (pyrometasomatism) produced by hydrothermal solutions generated during the latter stages of magmatic differentiation. Where batholiths have been intruded into rocks which are already regionally metamorphosed, the contact rocks formed are injection gneisses or migmatites. *See* AUREOLE, CONTACT.

Igneous rocks make room for themselves by forceful injection (dilatance), by engulfing wall rock blocks (magmatic stoping), or by subsidence of overlying rocks. The hypothesis of granitization maintains that granites result from the wholesale transformation of sedimentary or metamorphic rock layers by solutions operating through mineral replacement or by ionic emanations acting through solid diffusion. *See* GRANITIZATION.

Blocks of wall rock included in an intrusive mass are xenoliths; their partial destruction by reaction may produce irregular clumps of mafic minerals called schlieren. In some instances such endomorphic effects are sufficiently intensive to result in modification of the composition of the magma (syntaxis). *See* XENOLITH.

Crystallizing under equilibrium conditions, early magmatic minerals react with remaining fluid to yield new species (see diagram). Interruption of the sequence will yield liquid fractions richer in silicon dioxide, alkalies, iron, and water than the original magma and crystalline fractions richer in calcium and magnesium than the parent magma (magmatic differentiation).

Igneous rocks occur in clans or associations which possess characteristic trace elements and appear in specific structural provinces (Table 3). The origins of various igneous rocks are summarized in Table 4. *See* PETROGRAPHIC PROVINCE.

**Sedimentary rocks.** With the exception of material deposited by glaciers (till; or the consolidated form tillite), sedimentary rocks show bedding or stratification. This separation into generally parallel layers (beds, strata) results from sorting according to grain size during deposition, from

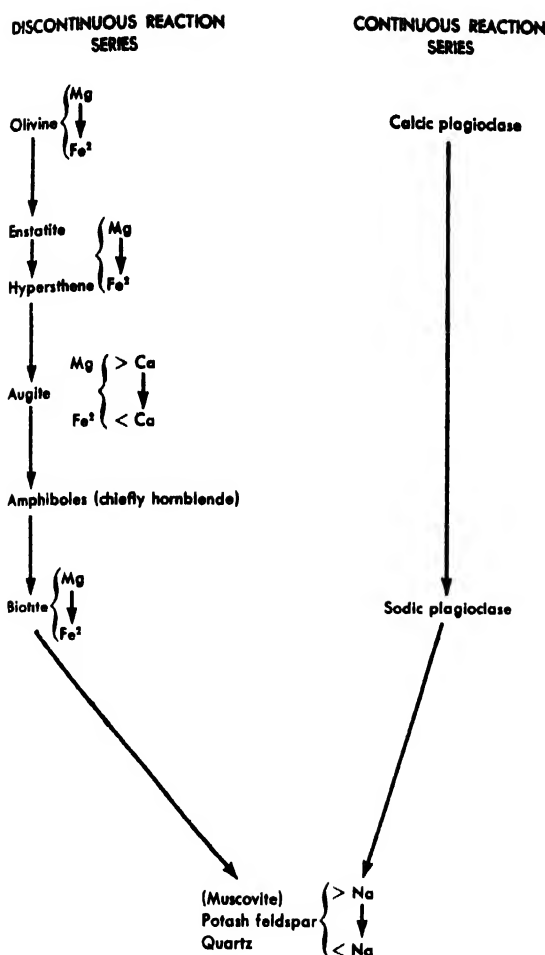
Table 1. Types of volcanic structure

Name	Characteristics
Shield	Low height, broad area; formed by successive fluid flows accumulating around a single, central vent
Cinder cone	Cone of moderate size with apex truncated; circular in plan, gently sloping sides; composed of pyroclastic particles, usually poorly consolidated
Spatter cone	Small steep-sided cone with well-defined crater composed of pyroclastic particles, well consolidated (agglomerate)
Stratocone	Composed of interlayered flows and pyroclastics; flows from sides (flank flows) common, as are radial dike swarms, slightly concave in profile, with central crater
Caldera	Basins of great size but relatively shallow; formed by explosive decapitation of stratocones, collapse into underlying magma chamber, or both
Plug-dome	Domal piles of viscous (usually rhyolitic) lava, growing by subsurface accretion and accompanied by outer fragmentation
Cryptovolcanic structures	Circular areas of highly fractured rocks in regions generally free of other structural disturbances; believed to have formed either by subsurface explosions or sinking of cylindrical rock masses over magma chambers

Table 2. Characteristics of intrusive igneous rock masses

Name	Shape	Structural relations to wall rocks	Size and other features
Dikes ✓	Tabular, lensoid	Discordant	Few feet to hundreds of miles long
Sills	Tabular, lensoid	Concordant	Up to several hundred feet thick
Laccoliths	Plano-convex or doubly convex lenses	Generally concordant	1-4 mile dia several thousand feet thick
Volcanic necks	Pipelike	Discordant	Few hundred feet to a mile in diameter, cores of eroded volcanoes
Stocks	Irregular, with steep walls	Crosscutting	A small batholith or its upward projection; outcrop area less than 40 mi <sup>2</sup>
Batholiths	Irregular, contacts dip steeply or outward; no bottoms known	(1) Discordant (2) Concordant in general, may be cross-cutting in detail	Some cover 16,000 mi <sup>2</sup> ; some are composite intrusives of varied petrology
Plutons	Irregular	Usually cross-cutting	Usually large; used as general name for intrusive masses that do not fit other definitions





Reaction series of Bowen (modified).

differences in composition or texture, or from variations in the rate of deposition. The development of most sedimentary rocks proceeds in the following stages: (1) There is a source rock, any older rock or, for organic sediments, a supply of organically originated material; (2) By weathering, the older rock is mechanically comminuted, chemically altered, or both, to form unconsolidated surficial rock debris called mantle; (3) Particles are transported by streams, ocean and lake currents, wind, glaciers, or by the direct action of gravity which causes particles to slide and roll down slopes; (4) Material moved by rolling, suspension, or solution is deposited; and (5) Deposits usually are consolidated by the processes of cementation (sandstones), compaction (shales), and recrystallization (limestones). Chemical changes accompanying consolidation are termed diagenetic. Weathered material not transported may become a residual sedimentary rock (bauxite). Sedimentary rocks are deposited either on land areas (continental) or in ocean waters (marine). Most marine sedimentation takes place on the submarine extensions of the continents called continental shelves. Examples of types of sedimentary deposits are listed in Table 5. Features characteristically found

in sedimentary rocks, in addition to stratification, are crossbedding, concretions, ripple marks, mud cracks, and fossils. See DIAGENESIS; SEDIMENTATION (GEOLOGY); WEATHERING PROCESSES.

A formation, which is the basic unit of stratigraphy, is a series of rocks deposited during a specific unit of geologic time and consisting either of a particular rock type or of several types deposited in a sedimentary cycle. Such a cycle is the changing sequence of deposits reflecting, for example, advance or retreat of marine waters in a particular area. However, while sandstone may be deposited at one time in one place in the sedimentary basin, limestone may be formed simultaneously elsewhere. Such lateral variation in a formation is referred to as facies. See CYCLOTHEM; FACIES (GEOLOGY); STRATIGRAPHY.

By means of detailed studies of the fossils of a formation and its lithology, composition, structure, and distribution, the paleoecology of the area may be reconstructed. Correlation of formations is attempted chiefly on the basis of fossils, with supplementary data from the lithology, stratigraphic position, insoluble residues (in acid-soluble rocks), heavy detrital minerals (in clastic rocks), and in

Table 3. Igneous rock clans or associations

Name	Main rock types	Environment
Oceanic Olivine basaltic	Olivine basalt, minor trachyte, peridotite	Volcanic islands of deep oceanic basins
Alkaline volcanic	(a) Olivine basalt, trachyte, phonolite (b) Leucite basalt, trachybasalt, trachyte (c) Spillite, khataphyre	(a, b) Nonorogenic continental regions (c) Orogenic continental regions; former geosynclines
Tholeiitic basaltic	Basalt (generally olivine-free), quartz diabase	Continental plateau areas
Calc-alkalic volcanic Lopolithic	Andesite, rhyolite, basalt Norite, gabbro, anorthosite, peridotite	Continental orogenic areas Lopoliths, thick differentiated sheets
Alpine-type peridotites Precambrian anorthositic	Peridotites, serpentinites Andesine or labradorite anorthosite, norite, syenite, monzonite	Orogenic zones Domed pluton of massifs in Precambrian terranes
Granite batholithic	(a) Simple: granite, granodiorite (b) Complex: gabbro, tonalite, granodiorite, minor granite	Precambrian shields; cores of mountain ranges
Minor granitic intrusive	Granite (some alkalic), quartz syenite, syenite, diorite	Hypabyssal, in mountain ranges and as their outliers
Nepheline syenitic	Feldspathoidal rocks, carbonatites	(1) Simple plutons (2) Ring complexes
Lamprophyric	Minette, kersantite, camptonite	Dike swarms

drill holes by electrical conductivity, radioactivity, and seismic wave velocities.

**Metamorphic rocks.** Metamorphism transforms rocks through combinations of the factors of heat, hydrostatic pressure (load), stress (directed pressure), and solutions. Most of the changes are in

texture or mineral composition; major changes in chemical composition are called metasomatism. The major types of metamorphism are presented in Table 6. Rocks that can serve as parent material for metamorphic derivatives include both igneous and sedimentary types, and, less commonly, older

Table 4. Synopsis of magmatic evolution\*

Primary magmas	Mode of origin	Common types of igneous rocks
Alkaline olivine basalt magmas	By differentiation and plutonic crystallization	Teschenites Theralites Essexites Shonkinites Pyroxenites Picrites Peridotites { Syenites Nepheline syenites Ijolites → Carbonatites
	By differentiation and volcanic crystallization	Mugearites Trachybasalts Olivine basalts Picrite basalts { Alkali rhyolites Trachytes Phonolites
	By contamination involving reaction with "granite" of continental basement	Nepheline basalts Soda lamprophyres Melilite basalts { Phonolites
	By differentiation and modification in geosynclines	Keratophyres Spilites { Leucite basalts Biotite lamprophyres { Leucite phonolites Latites Trachytes
	By differentiation and plutonic crystallization	Granites Granodiorites Tonalites Diorites Syenites { Pegmatites Aplites
Granodiorite granite magmas	By differentiation and volcanic crystallization	Rhyolites Dacites Andesites Latites { Dacites Rhyolites
	By mixing with basaltic magma	Andesites
	By assimilative reaction with slates, limestones, amphibolites, etc.	Syenites Nepheline syenites { Diorites Tonalites
	By differentiation and plutonic crystallization	Granophyres Diorites Anorthosites (Bytownite) Gabbros Pyroxenites Peridotites
Tholeiitic magmas	By differentiation in thick sills	Granophyres Quartz diabases Diabases Olivine diabases
	By differentiation and volcanic crystallization	Tholeiitic basalts Picrite basalts { Rhyolites Andesites

\* F. J. Turner and J. Verhoogen, *Igneous and Metamorphic Petrology*, 2d ed., McGraw-Hill, 1960.

metamorphic rocks as well. The complexity of the possible metamorphic mineral assemblages stems not only from the variety of possible parent rocks and from the imposition of the several kinds of metamorphism but also from variation in the intensity of particular types of metamorphism (grade), and from the difficulty of readily achieving chemical equilibrium through solid-state reactions. Various features characteristic of metamorphic rocks include foliation (slaty cleavage, schistosity, and gneissic structure), lineation, banding, and relief structures. See METAMORPHISM; METASOMATISM.

The facies principle is employed in attempting to reconstruct the environment under which a metamorphic rock was developed. A metamorphic facies consists of all rocks, without respect to chemical composition, that have been recrystallized under equilibrium within a particular environment of stress, temperature, load, and solutions. The first two factors are considered critical. The facies are named after metamorphic rocks deemed diagnostic of such restricted conditions. In practice one assigns a group of related rocks of different compositions to a particular facies upon the presence of such a key assemblage.

A. Facies of contact metamorphism. Load pressure low, generally 100–3000 bars. Water pressure highly variable, in some cases possibly exceeding load pressure, in a few cases very low. Facies listed in order of increasing temperature for given range of pressure conditions.

1. Albite-epidote hornfels (formerly albite-epidote amphibolite facies, actinolite-epidote hornfels subfacies)
2. Hornblende hornfels (formerly amphibolite facies, cordierite-anthophyllite subfacies)
3. Pyroxene hornfels
4. Sanidinite—corresponds to minimum pressures (load,  $P_{H_2O}$ ,  $P_{CO_2}$ ) and maximum temperatures—pyrometamorphism.

B. Facies of regional metamorphism. Load and water pressures generally equal, high (3000–12,000 bars). Facies listed in order of increasing temperature and pressure.

1. Zeolitic (hitherto not recognized)
2. Greenschist
  - a. Quartz-albite-muscovite-chlorite (formerly muscovite-chlorite)
  - b. Quartz-albite-biotite (formerly biotite-muscovite)
  - c. Quartz-albite-almandine (formerly albite-epidote amphibolite facies, chloritoid almandine subfacies)
3. Glaucophane schist (hitherto of uncertain status; previously equated with the greenschist facies; the glaucophane schists and their associates seem to represent a divergent line of metamorphism conditioned by development of unusually high pressures at low temperatures)

4. Almandine amphibolite
  - a. Staurolite-quartz
  - b. Kyanite-muscovite-quartz
  - c. Sillimanite-almandine
 (formerly staurolite-kyanite)
5. Granulite
  - a. Hornblende granulite
  - b. Pyroxene granulite
6. Eclogite

Regional variations in grade may be mapped by means of isograds, lines formed by the intersection

**Table 5. Selected examples of sedimentary deposits under various environments**

Agent	Deposit	Resulting rock
<b>Continental</b>		
Streams	Valley fill	Sandstone
	Alluvial fan	Conglomerate
	Delta	Siltstone
Lakes	Varved clay	Shale
Springs		Travertine
		Silicious sinter
Swamps	Peat	Coal
Wind	Dune	Sandstone
	Dust	Loess
	Volcanic ash	Tuff
	Moraine	Tillite
Glaciers	Stalactite	Dripstone
Groundwater	Talus	Breccia
Gravity	Avalanche	Conglomerate
	Landslide	
<b>Marine</b>		
Breakers and alongshore currents	Beach	Sandstone
		Conglomerate
Longshore currents		Sandstone
Marine organisms	Reefs and other shell deposits	Shale
		Shell limestone
		Coquina
Marine water	Evaporites	Diatomite
		Rock salt
Marine water	Colloidal precipitates	Rock anhydrite
		Phosphorite
		Manganese oxide concretions
		Chert

**Table 6. Types of metamorphism, their factors and results**

Type	Factors	Changes in rock
Cataclastic	Stress, low hydrostatic pressure	Fragmentation, granulation
Contact (thermal)	Heat, low to moderate hydrostatic pressure	Recrystallization to new minerals or coarser grains; rarely melting
Pyrometamorphism	Heat, additive hydrothermal solutions, low to moderate hydrostatic pressure	Reconstitution to new minerals; change in rock composition
Regional (dynamic)	Heat, weak to strong stress, moderate to high hydrostatic pressure, $\pm$ nonadditive solutions	Recrystallization to new minerals or coarser grains; parallel orientation of minerals to produce foliation

of planes of isometamorphic intensity with the earth's surface. These are defined on the appearance of a specific mineral known to reflect a major increase in the intensity of metamorphism.

The primary cause of stresses acting during regional metamorphism is diastrophism of the mountain-building type. The higher temperatures may result from deep burial, owing to the geothermal gradient of the earth, in part to concentrations of radiogenic heat, or in part to heat supplied by cooling masses of magma. In contact metamorphism this last is the sole heat source. See GEOLOGIC THERMOMETRY.

Once formed, metamorphic rocks are subject to further changes through folding and crumpling of the foliation and through extensive injection of igneous material to form migmatites. [E.W.H.]

**Bibliography:** T. F. W. Barth, *Theoretical Petrology*, 1952; W. S. Fyfe, F. J. Turner, and J. Verhoogen, *Metamorphic reactions and metamorphic facies*, *Geol. Soc. Am. Memoir* 73:259, 1958; E. W. Heinrich, *Microscopic Petrography*, 1956; P. Niggli, *Rocks and Mineral Deposits*, 1954; F. J. Turner and J. Verhoogen, *Igneous and Metamorphic Petrology*, 2d ed., 1960.

## **Petromyzontiformes**

One of two orders of Recent jawless fishes containing the lampreys. This order is also known as the Petromyzontia. They are degenerate, modern representatives of the plated cephalaspidomorphs of Silurian and Devonian times. Lampreys differ fundamentally from the superficially similar but remotely related order Myxiniiformes or hagfishes. The snout is produced and overlies a circular oral disk that bears cornified teeth and lacks enlarged barbels; the single nostril is located on top of the head before the eyes and does not penetrate the palate; there are seven pairs of spherical gill pouches that open internally into a respiratory tube and externally through seven pairs of pores; the one or two dorsal fins are more or less distinct from the caudal fin; and there are two pairs of semi-circular canals. Lampreys are bisexual animals that live permanently in fresh water or enter streams to breed. Eggs are deposited in gravel riffles, and the young develop into blind larvae that burrow in the soft bottom and feed for several years on micro-organisms strained from the water. At metamorphosis, the oral disk and eyes develop and the transformed lampreys become free swimming.



Sea lamprey, *Petromyzon marinus*. (After G. B. Goode, *Great International Fisheries Exhibition, London, 1883*, U.S. Natl. Museum Bull. 27)

Some species, known as brook lampreys, remain in streams for a few months without feeding, then breed and die. Most lampreys, however, are parasitic, feeding on other fishes and growing for a year or more before they breed and die. A protrusile tongue armed with horny teeth is employed to rasp an opening in the host's body, from which blood is sucked. So destructive are these parasites that, after gaining recent access to the upper Great Lakes, the sea lamprey (*Petromyzon marinus*) all but exterminated the lake trout and other valuable commercial fishes.

Lampreys are classified in a single family, Petromyzontidae, with 7 genera and about 30 species. They are chiefly inhabitants of temperate and cold-temperate waters in both the Northern and Southern Hemispheres. See CYCLOSTOMATA (CHORDATA)

[R.M.B.]

## **Pewee**

A name usually applied to the wood pewee, *Contopus virens*, a moderate-sized flycatcher of the family Tyrannidae. This dusky forest flycatcher can be distinguished from its relatives by the absence of an eye ring and the presence of two white wing bars. Its clear, flutelike call, pee-a-wee.



The wood pewee, *Contopus virens*; length to 6¼ in. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

with the middle syllable the low tone, is a characteristic song of the eastern deciduous forest, especially as the first song of the morning. It is replaced in the western forests by the western wood pewee, *C. sordidulus*. The name pewee is also often applied to the phoebe. See FLYCATCHER; PASSERIFORMES; PHOEBE. [J.D.B.]

## **Pewter**

An alloy containing tin and lead, usually in proportion of four to six parts of tin to one of lead. Other metals are sometimes used with or in place of the lead, including copper, antimony, and zinc. Pewter has limited use in the making of ornamental wares;

until the development of cheap china early in the nineteenth century, pewter was very commonly used throughout Europe and America for decanters, mugs, tankards, bowls, dishes, candlesticks, and canisters. The finer grades, those containing the most tin, were also used in making religious and civic vessels such as communion plates, chalices, and cruets, and symbolic cups and flagons. See TIN. [G.CO.]

## pH

A term used to describe the hydrogen-ion activity of a system. It is defined by the expression  $\text{pH} = -\log_{10} a_{\text{H}^+}$  in which  $a_{\text{H}^+}$  is the activity of the hydrogen ion. In dilute solutions, the activity is essentially equal to the concentration, and the pH may be approximately defined as  $\text{pH} = -\log_{10} [\text{H}^+]$ , where  $[\text{H}^+]$  is the concentration of the hydrogen ion in moles per liter. The use of the pH makes negative exponents unnecessary in describing the hydrogen-ion activity. In a system in which the hydrogen-ion activity is  $10^{-3}$  mole/liter, the pH is 3. The term is seldom used to describe solutions in which the hydrogen-ion activity is 1 or greater.

The expression defining the pH is closely related to the free energy of the hydrogen ion with respect to a standard reference state. See ACID AND BASE; GLASS ELECTRODE; HYDROGEN ELECTRODE; HYDROGEN ION. [F.J.J.]

**Bibliography:** R. G. Bates, *Electrometric pH Determinations*, 1957.

## Phaeophyta

These plants, also called brown algae, are so characteristically marine that the larger members are commonly known as brown seaweeds or kelps. They are widely distributed but are most numerous on the rocky coasts of the colder oceans. However, they may occur in the open ocean, on muddy salt marshes, or as epiphytes or endophytes. Many are able to withstand the desiccation of the intertidal zone, whereas others must remain continuously submerged. A few genera may be found at considerable depth. Both annual and perennial species have been discovered. They are distinguished from other algae by their brown or olive-green color, by their reproduction, and by the structure of the plant body.

The brown color is caused by a marked predominance of xanthophylls, including fucoxanthin, in the yellowish-brown chromatophores, and these pigments mask the green of the chlorophyll (see CAROTENOID; CHLOROPHYLL). Carbohydrates are stored not as starch but as soluble sugars, mostly in the form of the polysaccharide, laminarin, and the alcohol mannitol.

**Economic importance.** Although of negligible importance in Europe and in the Americas, several of the brown algae are utilized as human food in the Orient, chiefly Japan and the lands surrounding the China Sea. Here *Laminaria*, *Alaria*, and other kelps are used in a variety of ways. They may be used raw or cooked, or they may be dried and pack-

aged for flavoring sauces, thickening soups, and for making wafers, cakes, or other confections. During the war years, the kelps were gathered and processed to produce organic components called alginates (salts of alginic acid obtained from the colloidal gel, algin, contained in kelps). Alginates are widely used in the ice cream, baking, rubber, and paint industries. The alginates are also used as suspending agents in a wide variety of pharmaceuticals.

**Reproduction.** All of the Phaeophyta have an alternation of generations, that is, a definite alternation of sexual and asexual individuals. The plant which develops the gametes (sex cells) is called the gametophyte, and that which produces the asexual spores is termed the sporophyte. The gametophyte of this group may be small, filamentous, and inconspicuous, or it may not differ markedly from the sporophyte. In such forms as the Fucales the gametophyte stage is obscurely represented by a one-celled reproductive agent, the gamete. This evolving pattern of alternation of generations within the algae is considered important since it is a characteristic of all higher plants.

Pear-shaped motile reproductive cells with two lateral flagella are characteristic of the phylum. Sexual reproduction is isogamous or anisogamous involving a small, active male gamete and a large, nonmotile egg. Asexual reproduction is accomplished by fragmentation of the thallus or in most species by zoospores (swimming spores).

**Morphology.** The plant body is always multicellular and may vary from a simple, filamentous

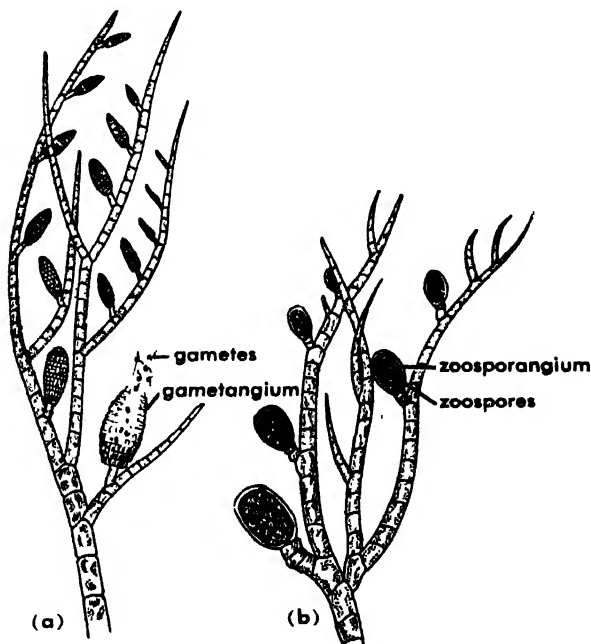


Fig. 1. *Ectocarpus*. (a) Branched filament bearing multicellular gametangia. One gametangium is releasing several biflagellate, pear-shaped gametes. (b) Branched filament bearing unicellular zoosporangia with zoospores. (From H. J. Fuller and O. Tippo, *College Botany*, rev. ed., Holt, 1954)

form to a complex body having a basal attachment structure and a stemlike main plant body. The latter may be branched or unbranched and of lengths varying up to about 50 meters. They may be broad and ribbonlike or have leaflike branches, and are often provided with air bladders. The largest forms may have a relatively complex internal structure.

**Classification.** The Phaeophyta are divided into three classes on the basis of their life cycles. The Isogeneratae are distinguished by having an isomorphic alternation of generations. A typical example of this group is *Ectocarpus* (Fig. 1), a genus of world-wide distribution with many species. It grows upon rocks or is epiphytic upon plants, usually other brown algae. The plant body is composed of branched filaments which may be 5-6 in. long. Each cell is uninucleate with one or more simple or lobed plastids. The plant body may be either gametophytic or sporophytic (isomorphic). The diploid zygote is the first cell of the sporophytic generation which develops into the usual branched plant.

Two kinds of reproductive organs may be produced by the sporophyte. The terminal part of certain branchlets enlarges forming a sporangium. The single nucleus of the young sporangium divides meiotically and then the daughter cells subdivide mitotically until there are 32 or 64 free nuclei (see MEIOSIS; MITOSIS). There ensues a cleavage of the protoplast into small uninucleate protoplasts, each containing a single chromatophore. Each of these protoplasts develops into a pyriform (pear-shaped), laterally biflagellate zoospore. These zoospores are produced within a common cavity which opens through an apical pore; this type of sporangium is unilocular. Meiosis occurs in the first of the nuclear divisions, so that the zoospores and the resulting plants are haploid. The terminal portions of other lateral branches, by numerous vertical and transverse cell divisions, may produce an aggregation of small cubical cells each containing a protoplast which may become a zoospore. Each of these zoospores is formed from a separate cell; such sporangia are said to be plurilocular. The zoospores are diploid and upon germination produce additional sporophytes. These gametophytic plants (developed from zoospores formed in the unilocular sporangia) produce only plurilocular sporangia. The motile cells formed in sporangia may unite in pairs to form zygotes, the beginning of a new diploid generation, or may develop asexually into plants which are haploid and gametophytic.

Neither sexuality nor an alternation of generations is obligate, as both the gametophyte and the sporophyte can be reproduced asexually. Other examples of common genera belonging to this order are *Cutleria* and *Dictyota*.

The second class, the Heterogeneratae, is distinguished by a heteromorphic alternation of generations. This includes the order Laminariales, the species of which are commonly known as the kelps. In *Laminaria* the sporophyte is a large, complex,

perennial plant consisting of a branching holdfast, a stipe (stemlike structure), and an expanded blade (Fig. 2). This is the dominant generation. The gametophytes are microscopic branching filaments which, although free-living, are relatively ephemeral. Male and female gametes are produced on separate gametophytes, and gametic union is oogamous. It is of interest that the relation of the two generations is very similar to that found in the higher ferns and related groups.

The third class, Cyclosporeae, is a group in which there is only a free-living diploid generation. There is but one order, the Fucales, which contains such well known genera as *Fucus*, *Ascophyllum*, and *Sargassum*. In *Fucus* (Fig. 3), the production of reproductive cells is confined to the tips of the dichotomously branched thallus. Egg and sperm may be produced on the same plant or on different plants, depending on the species. As in all Phaeophyta, the sperm are laterally biflagellate. The

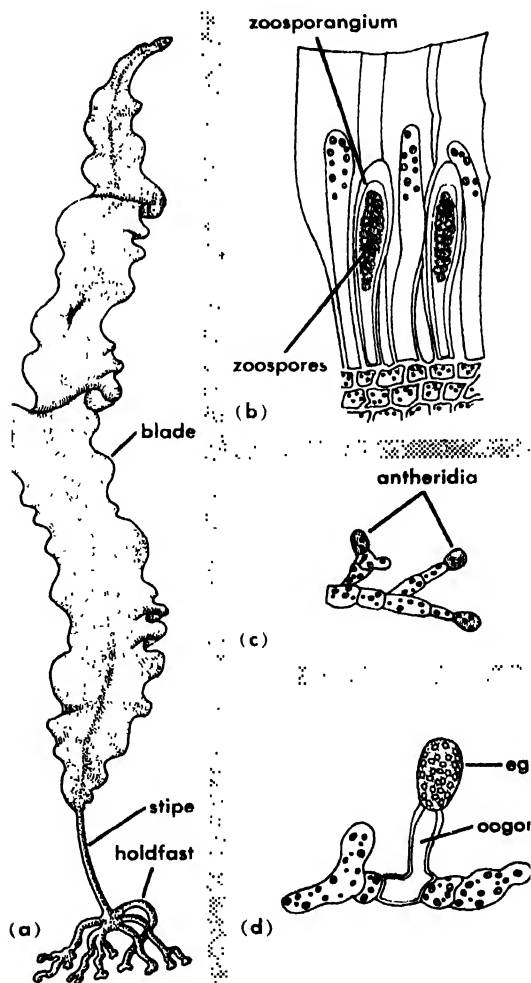


Fig. 2. *Laminaria*. (a) Devil's apron, with root holdfast, stipe, and blade. (b) Section of surface blade showing two zoosporangia with many zoospores. (c) Male gametophyte with terminal antheridia. (d) Female gametophyte with oogonium from which an egg is emerging. (From H. J. Fuller and Tippo, *College Botany*, rev. ed., Holt, 1954)

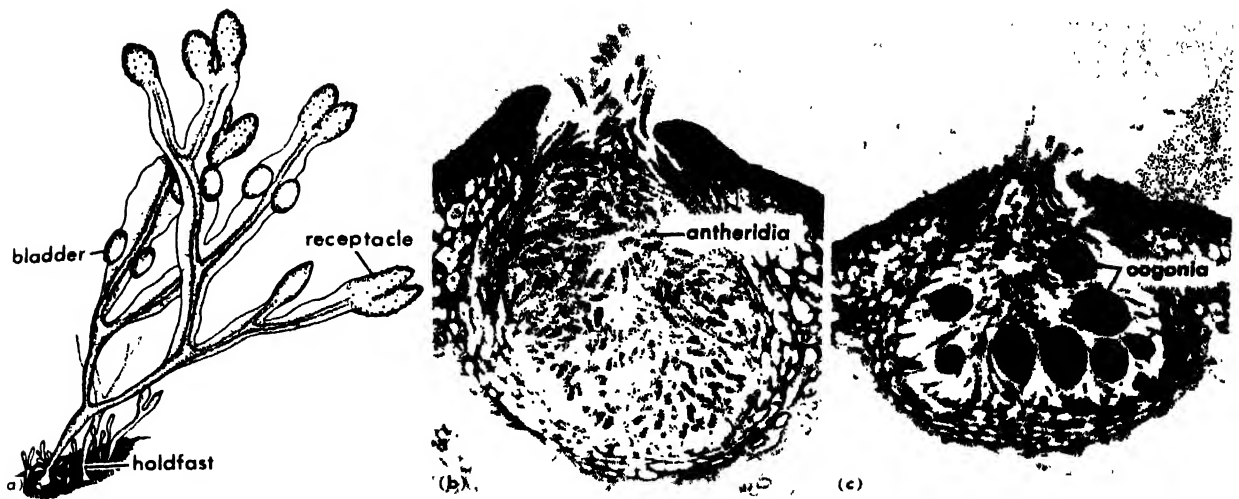


Fig. 3. *Fucus*. (a) Plant with basal disk or holdfast, dichotomous thallus, paired bladders or floats, and inflated tips or receptacles. The black spots on the latter are the openings leading into the cavities or conceptacles. (b) Section through a male conceptacle

lined with paraphyses (multicellular hairs) bearing antheridia. (c) Section through female conceptacle lined with paraphyses and with a few oogonia containing eggs. (From H. J. Fuller and O. Tippo, *College Botany*, rev. ed., Holt, 1954)

eggs produced in an oogonium vary from eight in *Fucus* to one in *Sargassum*. A meiotic division takes place in the formation of the haploid egg and sperm from the diploid plant. After the liberation of the gametes when nuclear union (fertilization) occurs, the diploid zygote thus formed germinates almost immediately. See GENETICS. [P.A.V.]

Bibliography: See THALLOPHYTES.

## Phagocytosis

A term used in general biology to refer to the engulfment of a particle by a cell, a process important in the nutrition of many protozoa and primitive metazoa, and in the metamorphosis and disposal of metabolic products of higher forms. In medicine, phagocytosis refers especially to the engulfment of invading microorganisms by the wandering polymorphonuclear and mononuclear cells or microphages of the blood and the macrophages of the reticuloendothelial system, including the wandering histiocytes and the various fixed phagocytic cells of the spleen, liver, lymphoid tissue, and bone marrow. Phagocytosis is thus one of the primary defense mechanisms in immunity against infection. See CARDIOVASCULAR SYSTEM; METAZOA; PROTOZOA.

The degree of phagocytosis in any particular instance depends, first, on the nature of the microbial surface encountered. The virulence of many bacteria can be correlated with their possession of particular polysaccharide or protein surface components that inhibit phagocytosis. The surface properties can frequently be altered by contact with specific antibody—an opsonin, or antibody plus complement (opsonic action), so that the sensitized cell is now more readily phagocytized. A second factor is the nature of the tissue or other surface on which the microorganisms and phagocytic cells rest; rough surfaces promote an en-

hanced surface phagocytosis. In the body, bacteria may, in the presence of the appropriate antibody and complement, become adherent to particles such as erythrocytes (immune adherence), and in this condition, they are more susceptible to phagocytosis than when free.

A variety of consequences may ensue once the microorganisms have been engulfed. In many instances, bacteria are readily digested and destroyed; if the other factors in pathogenicity are favorable to the host, full protection and recovery follow. However, many important pathogens, such as the staphylococcus, can survive and often multiply within leukocytes. Intracellular microorganisms are characteristically found in tularemia, typhoid fever, brucellosis, and tuberculosis, among other chronic diseases. In this state, the bacteria may actually be protected against the bactericidal actions of antibody and complement, as well as some, although not all, antibiotics. The ultimate resolution of such infections depends on a complex balance between host and parasite. See ANTIBIOTIC; BRUCELLOSIS; IMMUNITY; OPSONIN; STAPHYLOCOCCUS; TUBERCULOSIS; TULAREMIA; TYPHOID FEVER; VIRULENCE. [H.P.T.]

Bibliography: R. A. Nelson, Jr., The immune-adherence phenomenon, *Proc. Roy. Soc. Med.*, 49:55-58, 1956; E. Suter, Interaction between phagocytes and pathogenic microorganisms, *Bacteriol. Rev.*, 20:94-132, 1956.

## Phalangida

An order of the class Arachnida. The members of this order, also known as the Opiliones, are characterized by an unsegmented cephalothorax broadly joined to a segmented abdomen, paired chelate chelicerae, paired palpi, four pairs of segmented legs, a pair of simple eyes, and a genital opening between the fourth coxae. Their bodies range from



less than 1 to over 15 mm in length. Respiration is by tracheae. Many species possess scent glands, whose openings are at the antero-lateral portion of the cephalothorax. These emit a material with a pungent odor. The phalangids lay eggs which hatch into forms resembling the adults. Their food consists of vegetable matter and soft-bodied insects.

While members of this order are found throughout the world, they attain their greatest abundance and diversity in the moist tropics. In these regions, they are often bizarrely spined and possess elaborate dorsal color patterns. In temperate areas, the species are drab and often have long, spindly legs. These are the forms popularly known as daddy-long-legs or harvestmen.

There are three suborders: the Cyphophthalmi, small, mitelike forms; the Laniatores, with flattened, often colorful, bodies, found chiefly in tropical areas, some species adapted to cave life; the Palpatores, including the long-legged forms found in temperate areas. Common genera are *Leiobunum* and *Phalangium*. See ARACHNIDA.

[C. J. GOODNIGHT]

## Phalarope

Any member of the family Phalaropidae, consisting of three monospecific genera. They are shorebirds, but differ from sandpipers in several respects. The toes are webbed basally and have lateral membranes, as in Wilson's phalarope, *Steganopus tricolor*, or with lobed margins as in the two pelagic, marine species. The female is the larger and more brilliantly colored. It is also the female that does the courting and selection of the nest site, while the male incubates the eggs.

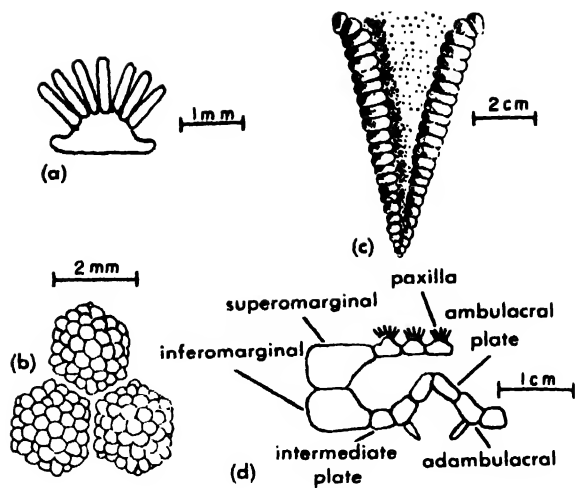


The Wilson's phalarope, *Steganopus tricolor*; length to 10 in. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

All three species nest in the Northern Hemisphere. Wilson's phalarope is found in the marshes of the Great Basin and Great Plains, whereas the two marine species are circumarctic. The northern phalarope, *Lobipes lobatus*, and the red phalarope, *Phalaropus fulicarius*, are both capable swimmers and may stay away from land for indefinite periods. See CHARADRIIFORMES; SANDPIPER. [J. D. BLACK]

## Phanerozonida

An order of Asteroidea in which pedicellariae may occur, but are not the crossed type, and in which the margins of the body are defined by two conspicuous series of marginal plates, one placed vertically above the other. The marginals constitute



Diagnostic features of Phanerozonida. (a) Paxilla, in side view. (b) Three paxillae in surface view (*Pseudarchaster*). (c) Arm of *Astropecten primigenius*, showing marginal plates and paxillae in transverse rows. (d) Transverse section of arm.

a buttressing skeleton which is usually more robust than the ambulacral skeleton. The upper surface is covered by symmetrical rows of plates which often bear brushlike clusters of spines, the paxillae. Papulae are restricted mainly to the upper surface. The tube-feet lie in two series in each ambulacral groove. Pentamerous symmetry is more constant than in the other orders of Asteroidea. The order is represented by Paleozoic fossils (suborder Pustulosa) and by three other extant suborders. See ASTEROIDEA; NOTOMYOTINA; PAXILLOSINA; VALVATINA.

[H. B. FELL]

## Pharetronida

An order of the subclass Calcinea in the class Calcarea. These sponges have a leuconoid structure. The main skeleton is composed of quadriradiates joined together by a calcareous cement or consists of a rigid calcareous network not composed of spicules. The dermal skeleton commonly includes spicules in the shape of tuning forks or may be composed of overlapping calcareous scales. Examples of this order are *Minchinella*, *Murrayona*, and *Petrobiona*. See CALCAREA; CALCARENEA.

[W. D. HARTMAN]

## Pharmaceutical chemistry

The chemistry of drugs and of medicinal and pharmaceutical products. The important aspects of pharmaceutical chemistry are

1. Isolation, purification, and characterization of medicinally active agents and materials from natural sources (mineral, vegetable, microbiological, or animal) used in treatment of disease and in compounding prescriptions

2. Synthesis of medicinal agents not known from natural sources, or the synthetic duplication, for reasons of economy, purity, or adequate supply, of substances first known from natural sources

3. Semisynthesis, so called, of drugs, whereby natural substances are transformed by comparatively simple steps into products with more favorable therapeutic or pharmaceutical properties

4. Determination of the derivative or form of a medicinal agent which exhibits optimum medicinal activity and at the same time lends itself to stable formulation and elegant dispensing

5. Determination of incompatibilities, chemical and biological, between the various ingredients of prescription

6. Establishment of safe and practical standards, with respect to both dosage and quality, in order to assure uniform and therapeutically reliable forms for all medication

7. Improvement and promotion of the use of chemical agents for prevention of illness, alleviation of pain, cure of disease, and search for new therapeutic agents, particularly where no satisfactory remedy now exists.

Chemistry, in its various facets, along with cognate disciplines from the physical and biological sciences, is essential for pharmaceutical chemistry. It is difficult to characterize any individual procedure or reaction as solely or preponderantly pharmaceutical. For example, the procedure for converting alcohol to ether, an excellent general aesthetic, may with slight modification be adapted for the production of ethylene, an important industrial chemical; and the step by which adiponitrile is hydrogenated to diaminoheptane, a synthetic fiber intermediate, is useful for the conversion of streptomycin to dihydrostreptomycin. Drug standardization uses many techniques and methods of analytical chemistry.

In the nineteenth century, beginning with the isolation of morphine from opium by F. W. A. Serturner (1805), an apothecary from Einbeck, Germany, effort was primarily directed toward the isolation of the active constituents of plants; this resulted in the discovery and description of many alkaloids, glycosides, carbohydrates, volatile oils, and fixed oils. The terms plant chemistry and pharmaceutical chemistry were used interchangeably, and this field of investigation remains active, even today, and promises to continue so for a long time to come, because it is estimated that less than 3% of the known flora have been examined.

The Food and Drug legislation of 1906 led to the broadening of research activities to include the development of analytical procedures for the maintenance of the quality and potency of drugs. These endeavors are continuing and expanding, as may be seen in the quinquennial revisions of the National Formulary and the U.S. Pharmacopeia, both official standards, as supplemented by state and federal regulations and tentative methods proposed by agencies charged with the enforcement of these official standards. See PHARMACOGNOSY; PHARMACOLOGY; PHARMACY. [W. H. HARTUNG]

**Bibliography:** G. L. Jenkins, W. H. Hartung, B. Data, and K. E. Hamlin, *The Chemistry of Organic Medicinal Products*, 4th ed., 1957; Na-

tional Formulary; Pharmacopeia of the United States.

## Pharmaceuticals testing

Methods used to determine whether pharmaceuticals are efficacious; of a standard and consistent potency and purity; and acceptable as to color, flavor, and physical appearance. Pharmaceuticals are medicinal products prescribed by medical doctors and dispensed through pharmacies and hospitals. Pharmaceuticals are taken into the body orally or through parenteral injection, usually when the patient is weakened by infection or illness.

The steps in the production cycle for pharmaceuticals must be uniformly controlled, and each operation must be as completely accurate as any other phase in the cycle. These control phases are commonly stated as (1) raw materials, (2) manufacturing procedures, (3) finished-product testing, and (4) control of identity.

**Raw materials.** These are usually referred to as "fine" chemicals and are purchased on specifications. If the raw material is officially recognized in the Pharmacopeia of the United States, or in the National Formulary, the specifications are provided in monographs in these compendiums. If the raw material is not officially included in these compendiums, chemical or physical specifications, or both, are prepared by the purchaser on the basis of the requirements for the finished product.

Physical specifications include such characteristics as bulk density, mesh size, color, odor, extraneous contamination such as fibers, and homogeneity. Chemical specifications usually include such characteristics as chemical or physiological potency, melting point, boiling range, optical rotation, moisture, heavy metals content, chemical identity, and presence of chemical contaminants.

Samples are taken upon receipt of a specific batch of raw material. Sampling may consist of a composite sample, composed of small portions from each container within the batch, or of one or more random small portions of the entire batch. The sample or samples are subjected to testing procedures to ensure conformance to each specification. In addition to physical inspection, testing procedures may range from simple melting-point determination to very complex chromatographic assays. Only after the raw material has been checked against each of the specifications can it be approved for use in pharmaceuticals.

**Manufacturing procedures.** To ensure products of the highest quality, pharmaceuticals must be manufactured under strictly regulated procedures and with adequate checks during each operation. Batch tickets or manufacturing formula cards set forth each manufacturing step in detail, and upon completion of each step, the card is usually initialed by the operator. Exact processing temperatures, specific mixing times, designated equipment, and precise details of operations, such as filtration or compression, are carefully spelled out on the batch ticket. All raw materials are double-checked

for identity and quantity before being incorporated in the process.

In-process assays are used to ensure homogeneity of mixing or completeness of reaction in the manufacturing process. Such assays range from simple pH measurements to complex infrared spectrophotometric determinations.

Upon completion of the manufacturing operation, batch tickets are usually checked by control inspectors to ensure that each step has been signed for. Representative samples of the bulk batch are taken by the inspectors and submitted to the chemical or biological testing laboratory for final assays. These representative samples may be either composite or random samples.

**Finished-product testing.** Usually each batch of a pharmaceutical must satisfy four requirements. It must conform to (1) the label claim for potency, (2) homogeneity standards, (3) standards of pharmaceutical elegance, and (4) identity requirements.

Potency standards require that the batch meet label claims within specified limits. Monographs of the United States Pharmacopeia or the National Formulary usually indicate maximum and minimum limits for official products. Limits for unofficial products are established by the manufacturer and are usually modeled after those for official products. Potency assays vary in complexity from a simple test on a single-component pharmaceutical such as an ascorbic acid tablet to very complex chemical and biological tests on a multicomponent pharmaceutical such as a vitamin preparation containing several vitamins plus minerals. Before approval, biological products must meet similar complex and severe criteria for potency.

Some special types of pharmaceuticals require additional complex tests. All parenteral products, intended for injection, must meet sterility requirements. Frequently tests are required on these for pyrogens and for "safety" (toxicity). These additional tests are necessary to ensure that no undesirable physiological reaction will result from administration of the pharmaceutical.

Statistical quality control trend charts on certain characteristics of a batch, such as tablet weights, ampule-filled volume, or random assay values, indicate conformance to homogeneity standards. Stability tests, which are usually more complex than potency tests, are frequently made to ensure that the pharmaceutical will remain potent and safe for use during normal shelf life. Such tests confirm the absence of harmful deterioration products during the ordinary time lapse between manufacture and use.

Pharmaceutical elegance refers to the physical appearance of the dosage units of pharmaceuticals. Conformance to these standards includes inspection to ensure that solutions are "sparkling clear," that tablets are not "capped" or "chipped," that parenteral ampules are free of "floaters," and that colored products are of the right hue or shade. These standards govern physical quality.

**Identity.** Identity is the final requirement in pharmaceutical testing. Testing for identity guar-

antees that the product has been properly labeled, that is, that the right product is in the right bottle with the right label. Serious consequences would result from a bottle of strychnine tablets carelessly labeled as saccharin tablets. To maintain the identity of the product, extensive checks are made throughout the manufacturing operation, including the use of duplicate label tags on all bulk goods, and very rigid controls are applied to printing, storage, and application of labels on finished pharmaceuticals to ensure final identity.

Only when all the operations in the production of pharmaceuticals, from securing raw materials to labeling the final container, are rigidly controlled through testing and checking procedures can one be assured that the pharmaceutical is pure, safe, and efficacious. See BIOASSAY; QUALITY CONTROL.

[W.B.F.O.]

**Bibliography:** W. B. Fortune, Control of fine chemicals and pharmaceuticals, *Anal. Chem.*, 29(7):17A-29A, 1957; *National Formulary*, vol. 10, 1955; *U.S. Pharmacopeia*, vol. 15, 1955.

## Pharmacognosy

The general biology, biochemistry, and economics of nonfood natural products of value in medicine, pharmacy, and other health professions. The products studied are of biologic origin, either plant or animal. They may consist of entire organs, mixtures obtained by exudation or extraction, or chemicals obtained by extraction and subsequent purification.

Pharmacognosy literally means knowledge of drugs, as do pharmacology and pharmacy. The center of interest in pharmacology, however, is on the mode of action of all drugs on the animal body, particularly on man. In pharmacy, major attention is directed toward provision of suitable dosage forms, their production and distribution. Pharmacognosy is restricted to natural products with attention centered on sources of drugs, plant and animal.

**Sources of materials.** Organs, or occasionally entire plants or animals, are dried or frozen for preservation and are termed crude drugs. They may be used medicinally in essentially this form, as in the case of the cardiac drug, digitalis, or the endocrine drug, thyroid, or as sources of mixtures or of chemicals obtained by processes of extraction.

Mixtures obtained by exudation from living plants include such drugs as opium, turpentine, and acacia. Processes of extraction are required to obtain such mixtures as peppermint oil (steam distillation), podophyllum resin (percolation), and parathyroid extract (solution). For a discussion of classes of natural products with medically significant members of this type see ESSENTIAL OILS; FAT AND OIL, EDIBLE; GUM; TERPENE; WAX, ANIMAL AND VEGETABLE.

Pure chemicals may be extracted from a crude drug (for example, the glycoside digitoxin from digitalis or the hormone insulin from pancreas), from a mixture obtained by exudation (for example, the alkaloid morphine from opium), or from

an extracted mixture (for example, the terpene menthol from peppermint oil). For a discussion of natural products of this type see ALKALOID; GLYCOSIDE; HORMONE.

Vitamins as a class of natural products are within the scope of pharmacognosy, although many are now obtained commercially by laboratory synthesis. Included also are antibiotics and biologicals (serums, vaccines, and diagnostic biological products). See ANTIBIOTIC; VITAMIN.

The general biology of pharmacognosy is largely descriptive. It includes the taxonomic position of the natural source of the product, the part of the plant or animal yielding the drug, the scientific and common names of the biologic source, the gross and histologic anatomic characterization of the part used, and the principal uses of the product in the health professions.

The biochemistry is both descriptive and experimental. It includes the chemical nature and percentage of the medically significant constituent, the mechanisms of biosynthesis of the constituent, and the role of the constituent in the economy of the plant. Increasing attention is being paid to mechanisms of biosynthesis by the use of radioactive precursors of medically active constituents to follow biosynthesis step by step. The isolation and chemical identification of new, potentially useful plant and animal constituents are an important aspect of biochemical research in pharmacognosy.

The economic aspects include the discovery and study of natural sources of crude drugs and their derivatives, development of cultivated sources where feasible, improvement of the yield of useful constituents, and protection of medically useful crop plants from their natural enemies. Methods of harvesting, drying, curing or other processing treatment, storing, packaging, and shipping enter into the commerce of drugs.

For a single drug, for example, menthol from peppermint oil, these several aspects are frequently inseparable. The commercial grower must know the species of *Mentha* yielding commercially profitable quantities of the essential oil and of menthol, percentage yields from various species, conditions suitable for growth, natural enemies of the growing plant such as specific viruses, methods of extraction of the oil and of isolation of the menthol, and proper conditions for packaging, storing, and shipping the purified drug. Scientific study of the genetics of species of *Mentha* are of academic interest and of potential commercial importance.

**Uses of materials.** Medical uses are chiefly as therapeutic, prophylactic, or diagnostic agents. Prior to the twentieth century, the materia medica of all countries was preeminently of natural products; it still is in less civilized countries. Twentieth-century research has contributed many synthetic and semisynthetic drugs to the modern materia medica, but a significant number of crude drugs are still the drugs of choice in therapy or serve as the source of widely used purified mixtures or chemicals. *Digitalis* and its glycosides, the alkaloids of opium and belladonna, penicillin, thy-

roid, insulin, and poliomyelitis vaccine are examples.

Pharmaceutical uses are chiefly in the production of palatable and stable dosage forms: gums and mucilages in emulsions and suspensions, starch and lactose in tablets, sugar and essential oils in elixirs, oils and waxes in ointments. Many natural products of insignificant or questionable therapeutic value continue to be used in home remedies.

Uses in other health professions include antiseptics, protectives, and local anesthetics used by dentists; rodenticides, insecticides, and other pesticides used in the protection of the public health; and a variety of prophylactic and therapeutic agents used by veterinarians.

A large number of natural products of value in the health professions are used also in cosmetology (essential oils, gums, fats, and waxes), in the culinary arts (spices, essential oils, and condiments) and in industry (naval stores, mucilages, fats, and waxes).

The role of a medically active chemical produced and used by the animal organism is usually well understood. Physiologic function within the organism and therapeutic use by man are usually closely related, as in the case of pepsin, thyroid, or the sex hormones.

Corresponding knowledge of medically active plant constituents is almost nonexistent. The role of menthol in the economy of *Mentha piperita*, of digitoxin in *Digitalis purpurea*, of morphine in *Papaver somniferum*, or of reserpine in *Rauwolfia serpentina* is unknown. Through the centuries, however, mankind has discovered that certain plants relieve the symptoms of or cure certain diseases. With the discovery of alkaloids in the early nineteenth century and of the other major classes of medically active plant constituents, the chemicals responsible for therapeutic actions have been identified one by one, but not the function of these chemicals in their respective plant sources.

**Types of materials.** Classes of therapeutic agents have frequently been discovered by study of biosynthesized medicinal chemicals. Most such classes, in fact, have been developed from chemicals originally known from crude drugs or from their exudates or extractives. The first uses of opium as a narcotic and analgesic drug are lost in antiquity, but its position in the medical practice of the day has been primary for over 2000 years. Morphine was among the first alkaloids to be isolated, has been widely used for more than 150 years, and from its study has developed a class of analgesic drugs of wide application in medical practice.

An analogous pattern has given the modern classes of hypotensive drugs and tranquilizers. The Indian drug *rauwolfia* after centuries of use in folk medicine eventually found its way into scientific medical practice in the orient. Study of its chemical derivatives during the 1930s and 1940s revealed the presence of many alkaloids one of which, reserpine, was shown to be an effective anti-hypertensive agent. Subsequent therapeutic use of

reserpine and other rauwolfia products soon demonstrated the tranquilizing action. A large class of drugs having hypotensive action, tranquilizing effect, or both has been developed rapidly. Various species of the genus have been characterized morphologically, and intensive study has been undertaken of practical methods of culture.

A third major class of modern drugs of natural origin, the antibiotics, has been developed largely since the beginning of World War II. The prototype, penicillin, was discovered in part as a result of fortuitous accident, but the many other commercially available antibiotics have been developed as a result of carefully planned systematic search.

Not infrequently, the clue that has led to collection and scientific investigation of a crude drug as a possible source of medically significant constituents has been use of the drug by uncivilized peoples for a nonmedical, but to them desirable, purpose such as narcosis, or as a poison against wild animals or man. The use of opium as a narcotic by the laity undoubtedly preceded its medical use. Coca was chewed or sucked by the Indians of South America to increase endurance, was condemned by the Spanish who conquered the Incas, but was eventually introduced into medical practice in Europe. Discovery of the local anesthetic action of the alkaloid cocaine led to the development of a new and important class of therapeutic agents, the local anesthetics.

A second native drug from South America, curare, was first used by the Indians as an arrow poison for killing wild animals used as food. The neuromuscular paralysis caused in game by the drug suggested therapeutic use as a muscular relaxant. Studies of the plants yielding crude curares revealed several species of two principal genera, *Strychnos* and *Chondodendron*, as the main sources. A number of crystalline alkaloids was isolated from crude curares; eventually the alkaloid tubocurarine was identified as having the therapeutic potentialities suggested by the paralyzing action of the native drug. The botanical source was established as *Chondodendron tomentosum*.

Comparable studies of the African arrow poisons inee and kombe have added ouabain and strophanthin to the class of cardioactive glycosides of which digitoxin is the most widely used. The arrow poisons are prepared from African species of *Strophanthus* and are used by natives of both the eastern and western coasts.

The alkaloid physostigmine, also from an African poison and useful in the treatment of glaucoma, was discovered as a result of use of its plant source as a human poison in the trial by ordeal of those accused of offenses. The alkaloid is the toxic constituent of the seeds of *Physostigma venenosum*, which were fed to the accused. Toxic symptoms were taken as evidence of guilt; those who vomited the material were considered guiltless.

**Synthetic materials.** Development of synthetic drugs related chemically to the active constituent of a natural product has frequently followed inves-

tigation of primitive use of the natural product as drug or poison. The objective of such development is usually to produce a drug having fewer undesirable side effects while retaining the useful therapeutic action. Substitutes for morphine, reserpine, cocaine, tubocurarine, and physostigmine are among a host of synthetic drugs which accomplish the objective to a greater or less degree and whose discovery depended upon study of natural products.

Intermediates useful in the laboratory synthesis of drugs often exist as therapeutically inactive chemicals in natural products. Plants and animals biosynthesize many such compounds with chemical structures similar to but not identical with medically useful substances. A slight change in molecular configuration may yield a potent therapeutic agent. A simple example is pinene, a chemical abundant in turpentine oil and convertible by laboratory procedures into camphor. The resulting "synthetic" camphor is actually semisynthetic and possesses the therapeutic and most other properties of natural camphor.

An important class of natural intermediates are the steroids, widely distributed in both plants and animals. Some chemical variations are active physiologically and as drugs, for example, sex and adrenal cortical hormones. Natural sources, glands of domesticated animals used as food by man, are not available in quantities adequate to fulfill the drug needs for these products. Many plants contain steroids suitable as intermediates for sex hormones. Natural intermediates readily converted into adrenal cortical hormones are uncommon, and extensive search for such steroids has been made since the late 1940s. Field studies involve collection and identification of plants judged to be potentially good sources of steroids, preliminary extraction and determination of the presence or absence of these intermediates, and further collection, drying, and preserving of larger quantities of promising species.

Systematic screening of plants of reputed therapeutic value and indigenous to a country or other restricted geographic area is a costly and time-consuming procedure—a major reason it has been done for relatively few regions. Notwithstanding, such surveys give promise of uncovering new sources for known classes of drugs, adding to knowledge of such little-known classes as hallucinogens and anticarcinogenic drugs, developing entirely new classes of therapeutic agents, and providing profitable natural sources of intermediates useful in drug synthesis. See ANTIMICROBIAL AGENTS; BIOCHEMISTRY; PATHOLOGY; PHARMACEUTICAL CHEMISTRY; PHARMACOLOGY; PHARMACY; PLANT PHYSIOLOGY; TRANQUILIZER.

[R.A.D.]

**Bibliography:** American Association of Colleges of Pharmacy, *Teachers' Seminar on Pharmacognosy and Related Subjects*, 1953, 1959; E. P. Claus (ed.), *Gathercoal and Wirth Pharmacognosy*, 3d ed., 1956; R. Pratt and H. W. Youngken,

*Pharmacognosy*, 2d ed., 1956; H. W. Youngken, *Text-Book of Pharmacognosy*, 6th ed., 1950.

## Pharmacology

The science of detection and measurement of the effects of drugs or other chemicals on biological systems. The effects of chemicals may be beneficial (therapeutic) or harmful (toxic) when administered to man, other mammals, or other living systems. The pure chemicals or mixtures may be of natural origin (plant, animal, or mineral) or may be synthetic compounds.

The broad area covered may be conveniently divided into a number of categories: chemotherapy, the use of chemicals to destroy invading organisms such as bacteria and molds in or on the host; pharmacotherapy, the use of drugs to restore or replace normal function in various tissue cells, organs, or integrated units; pharmacodynamics, studies on the mechanism of action of drugs which may utilize physiological, biochemical, or electrical techniques; toxicology, the study of the poisonous effects of chemicals; psychopharmacology, the study of the effects of chemicals on the behavior of man or animals; biochemical pharmacology, the effects of chemicals on biochemical reactions in living systems, and the effects of these systems on the chemicals, that is, their metabolism; structure-activity relationship, relationship of biological activity to chemical structure and molecular properties; and clinical pharmacology, the study and evaluation of the effects of drugs in man. See CHEMOTHERAPY; PATHOLOGY; TOXICOLOGY.

The chemicals which have a beneficial effect in man (such as restoring function or behavior toward the normal state, relieving pain, destroying harmful invading organisms, or aiding in the diagnosis of disease) are called drugs; those chemicals which produce only harmful effects are called poisons. All drugs may be poisonous, however, if administered in large enough amounts. See POISON.

It is a fundamental property of drugs that their effects increase in intensity as the dose is increased. Determination of this relationship delineates the dose-response curve. This relationship can be usefully summarized by the dose required to produce a standardized effect and also by the slope of the curve. In many situations, the numerical expression of the dose required to produce 50% of the maximal response permits comparison of drugs having the same type of action. This  $ED_{50}$  (dose effective at the 50% level) is usually the standard-effect dose used to compare the potency of drugs having the same action, since it can be determined with greater precision than other doses producing a smaller or larger percentage of the maximal effect. Measurement of toxicity as manifested by lethal effects is expressed similarly as the  $LD_{50}$  (lethal dose 50). The relative safety of drugs is estimated by means of the therapeutic index which is the ratio of the  $LD_{50}$  to the  $ED_{50}$ . The higher the ratio, the safer the drug. See EFFECTIVE DOSE 50; LETHAL DOSE 50.

The necessity to use such terms as  $ED_{50}$  and  $LD_{50}$ , which are averages, implies that individuals have been found to differ in the amount of drug that each will require to produce the desired effect. It is common knowledge that various individuals tolerate widely varying quantities of ethyl alcohol. Individual differences occur in the response to all drugs, and what may be an effective dose in one person may not be effective in another. Individuals may also differ qualitatively in their response to a drug. Depending upon the effect under study, placebos (pills or other preparations containing starch, sugar, and coloring matter, but no pharmacologically active agent) may also produce a favorable response in some individuals. The administration of a placebo to a person with hypertension may result in a temporary fall in blood pressure, for example; and in several studies of agents used to relieve pain, as many as 35% of the individuals tested obtained relief of pain from the administration of the placebo. Observations such as these have pointed out the necessity of using a placebo control or a positive control (a drug of established activity) in experiments designed to evaluate a new drug. Because the attitude of the physician who administers a drug may also influence the way patients respond, this variable is controlled by use of the double-blind technique; that is, neither the individual nor the doctor knows which sample is drug and which is placebo.

The use of animals has been and is absolutely essential to progress in pharmacology and other medical sciences; however, different species may differ markedly in their response to drugs both quantitatively and qualitatively. Drugs which produce sedation in one species may produce excitement in another. Fortunately, most chemicals produce similar effects in many species, including man. Unfortunately, all drugs found effective in animals are not effective in man, presumably due to differences in metabolism or excretion or to unrecognized differences in the nature of the disease process or the physiological mechanism of maintenance of the same function in different species. Study of these differences has yielded much basic information on mechanism of actions of drugs and on physiological and biochemical differences among species. Future studies may be expected to contribute to the understanding of individual differences within species. See BIOASSAY; PHARMACEUTICAL CHEMISTRY; PHARMACOGNOSY; PHARMACY; PSYCHOPHARMACOLOGIC DRUGS. [C.J.K.]

*Bibliography:* V. A. Drill (ed.), *Pharmacology in Medicine*, 2d ed., 1958; L. S. Goodman and A. Gilman, *The Pharmacological Basis of Therapeutics*, 2d ed., 1955.

## Pharmacy

The health profession concerned with the discovery, development, production, and distribution of drugs. Drugs are substances (other than devices) used to diagnose, prevent, cure, or relieve the symptoms of disease. For relations to closely allied



fields *see* MEDICINE; PHARMACEUTICAL CHEMISTRY; PHARMACOGNOSY; PHARMACOLOGY.

**General pharmacy practice.** This part of the profession is carried on in exclusive prescription pharmacies, semiprofessional pharmacies, and drug-stores. It consists of compounding and dispensing drugs on order of the physician, dentist, or veterinarian; serving as consultant on drugs to the health professions and to the public; and selling other health supplies such as antiseptics, bandages, and home remedies. Combination of nonprofessional with professional activities is customary in the United States, however, and is commonly termed retail pharmacy.

**Hospital pharmacy.** In addition to qualities characteristic of general pharmacy practice, hospital pharmacy includes special administrative features, provision of drugs for nursing stations, manufacturing of pharmaceutical preparations, teaching of nurses and medical and pharmacy interns, service to the hospital committee on pharmacy and therapeutics, and the preparation and revision of a hospital formulary. The hospital pharmacist may have charge of investigational drugs, radioactive pharmaceuticals, medical and surgical sterile supplies, and gaseous drugs for inhalation therapy.

**Pharmaceutical research.** One type of research is in pharmaceutical chemistry, synthetic if the objective is to produce new and improved drugs by laboratory procedures, and analytic if the objective is to provide improved methods of assay for quality control of pharmaceutical production. Research may be in product development, aimed at provision of more palatable, stable, economical dosage forms. It may be on drugs from natural sources (pharmacognosy), or on the mode of action of drugs from any source (pharmacology). It draws heavily on investigations in organic chemistry, biochemistry, and microbiology.

**Manufacturing pharmacy.** The pharmaceutical industry is the mass-production medium of drugs and suitable dosage forms. It produces some of its own raw materials, but depends on the closely allied chemical industry for most of them. It sponsors and conducts extensive pharmaceutical research.

Pharmaceutical distribution and promotion are specialized activities within the pharmaceutical industry which are now commonly called pharmacy administration; this term also includes the commercial phases of retail pharmacy.

**Pharmaceutical jurisprudence.** This is a highly specialized area in pharmacy. General pharmacy practitioners and hospital pharmacists are subject to Federal and state laws on drugs and pharmacy practice. A board of pharmacy serves as the law enforcement agency and the examining and licensing body in each state; various Federal agencies administer the provisions of the several Federal laws governing pharmacy.

Other special fields of pharmacy include journalism, administration of national and state associations, provisions of modern drug standards.

**Pharmaceutical education.** From 1932 to 1960 the education of pharmacists consisted of collegiate studies for at least 4 years in basic science, mathematics, general education, and professional areas. In 1960 the minimal educational program was extended by 1 year to a total of 5 years of collegiate study. A flood of new and complex drugs, released at the rate of hundreds per year since the end of World War II, has made necessary for the pharmacist a strong background in the physical and biological sciences, highly specialized training in drugs, and the sociologic and humanistic understanding desirable in all professional people.

[R.A.D.]

*Bibliography:* R. A. Deno, T. D. Rowe, and D. C. Brodie, *The Profession of Pharmacy*, 1959; Health News Institute, *Facts about Pharmacy and Pharmaceuticals*, 1958; E. W. Martin and E. F. Cook (eds.), *Remington's Practice of Pharmacy*, 1956.

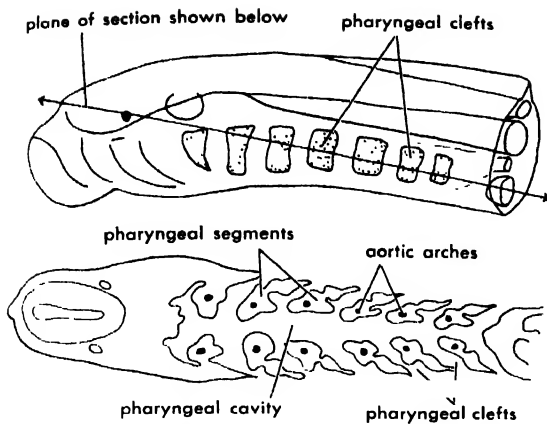
## Pharynx

A chamber at the oral end of the vertebrate alimentary canal, leading to the esophagus. Because of divergent specializations in the various classes of vertebrates, it cannot be described in general terms except at embryonic stages. In adult man it is divided anteriorly by the soft palate into a nasopharynx and an oropharynx, lying behind the tongue but anterior to the epiglottis; there is also a retropharyngeal compartment, posterior to both epiglottis and soft palate. The nasopharynx receives the nasal passages and communicates with the two middle ears through auditory tubes. The retropharynx leads to the esophagus and to the larynx, and the paths of breathing and swallowing cross within it. In adult fishes, the pharynx is not segregated from the mouth cavity but is pierced by a number of paired gill slits. *See* ESOPHAGUS; LARYNX.

**Embryology.** Shortly after the germ layers of the embryo are in place, the pharyngeal cavity appears as a simple enlargement of the anterior end of the endoderm tube. Its lateral or lateroventral walls promptly become thickened as a series of paired pillars, the pharyngeal segments (branchial arches or visceral arches), separated from their fellows by paired pouches which push outward from the endodermal lining of the cavity. These pharyngeal pouches are approached on each side by corresponding branchial grooves which push inward from the head ectoderm. The matching grooves and pouches may meet, their touching surfaces may then become thinned as closing plates, and they may actually break through as pharyngeal clefts. The mouth perforates into the pharyngeal cavity in a similar way anteroventrally. This embryonic appearance of the pharynx as an enlarged anterior gut chamber, whose walls are marked by pouches, grooves, clefts, and intervening pillarlike segments, is one of the few characteristics found in all members of the phylum.

The more primitive vertebrate embryos, such as those of lampreys and sharks, have up to seven or even more pairs of large open pharyngeal clefts of





Lamprey embryo, side view and horizontal section of pharynx region.

fairly uniform size. Embryos of reptiles, birds, and mammals show fewer segments and sharp reduction of the more posterior ones. Also several of the last pharyngeal pouches fail to break through as clefts. Whereas in most fishes several posterior clefts become enlarged and the intervening segments become equipped with gills and valves during development, few vestiges of either pouches or clefts remain at adult stages of land animals. Nevertheless, in all vertebrates the solid tissues of the pharyngeal segments give rise to numerous structures of the head and neck.

Each pharyngeal segment of the embryo is lined by ectoderm (the future epidermis) on the outside, and by endoderm (the future mucous membrane) on the inside. The mesoderm enclosed between these layers is of hypomeric or lateral-plate origin, and differentiates chiefly into muscles and arteries (see RESPIRATORY SYSTEM). In addition to these constituents, tongues of neural crest cells migrate down into the pharyngeal segments and differentiate there into skeletal elements, except in cyclostome fishes. See NEURAL CREST.

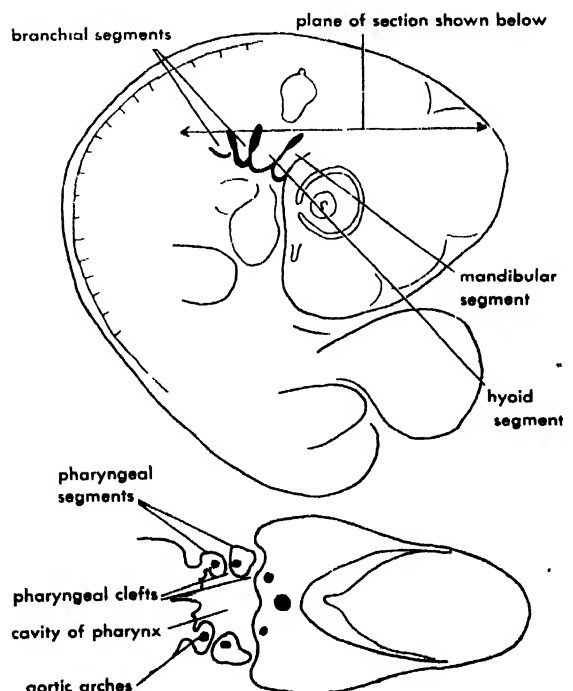
**Pharyngeal derivatives.** The first pair of pharyngeal segments, called the mandibular segments, enclose the mouth between them, and their skeleton-forming cells produce the mandibular cartilages, which are the earliest rudiments of the lower jaw. Their muscle-forming cells differentiate into a dorsal group of muscles whose principal function in all gnathostome vertebrates is to snap the jaws shut, and a ventral group which support the tongue by forming a sheet between the halves of the lower jaw. All these muscles are supplied by the motor division of the trigeminal nerve. See GNATHOSTOMATA.

The second pair are called the hyoid segments. They arise just posterior to the mandibular segments, usually separated from them temporarily by hyomandibular clefts. They lie ventral to the ear vesicles on the sides of the embryonic head. Their skeleton-forming cells form important parts of the tongue support, and their muscles are used in some vertebrates for opening the mouth, for tongue manipulation, for facial expression, and other func-

tions. The motor division of the facial nerve innervates the hyoid group of muscles. In higher vertebrates the hyomandibular pouch is involved in the formation of the auditory tube and the middle ear space.

The third and all more posterior pairs are called branchial segments since in fishes they usually form gill-bearing arches; they tend to be repressed or dispersed in the development of the higher vertebrates. Their skeleton-forming cells either form a succession of jointed rodlike bones or cartilages for the support of gills, as in fishes, or join the hyoid skeleton in tongue support and concentrate ventrally in the larynx cartilages, as in terrestrial vertebrates. In fishes the muscles derived from the first branchial segment (the third in the pharyngeal series) are strictly innervated by the glossopharyngeal nerve, and those of all the rest of the segments by individual branches of the vagus nerve. They function for manipulation of the gills, and for grinding and swallowing food. In higher vertebrates, the branchial segments become less and less distinct during development and the nerve supply, while still derived from the glossopharyngeal and vagus nerves, is not so clearly segmental, since the striated muscles themselves become arranged in a more or less continuous pharynx-constrictor sheet or congregate in the laryngeal cartilages, serving new mechanisms of swallowing and sound production. See SPEECH.

**Histology.** The pharynx is in general lined by simple mucous membrane, backed by fibrous connective tissue and a double layer of striated muscle. In bony fishes its skeletal arches may be thickly studded with simple teeth, or may even be devel-



Chick embryo, side view and horizontal section of pharynx region.

oped into grinding or crushing plates. Terrestrial vertebrates show simple tubular glands emptying, usually in great numbers, through the mucous membrane and tonsillar collections of lymphoid tissue in the submucosa. See TONSIL.

**Gross derivatives.** Elaborate gill pouches are developed in all aquatic groups and subject to many special adaptations. They differ sharply in design in lampreys, hagfishes, cartilaginous fishes, and the bony fishes. A median ventral evagination from the posterior end of the pharynx gives rise to the entire respiratory system of the land vertebrates, including lungs, larynx, and trachea. A similar but often dorsal evagination, usually from the pharynx-esophagus boundary, gives rise to the air bladder in bony fishes. See RESPIRATORY SYSTEM; SWIM BLADDER.

The epithelium of the pharyngeal pouches and of the pharynx floor produces a constellation of endocrine glands and other structures. See PARATHYROID GLAND; THYMUS GLAND; THYROID GLAND; ULTIMOBRANCHIAL BODIES. [W.W.B.]

**Bibliography:** L. B. Arey, *Developmental Anatomy*, 6th ed., 1954; A. S. Romer, *The Vertebrate Body*, 2d ed., 1955.

## Pharynx disorders

These include the more common congenital defects, inflammations, tumors, and nervous disorders which affect the pharynx.

Congenital defects commonly seen are malformed or split uvulae, or soft palates, and extension of a cleft palate backward to the pharyngeal region. See HARELIP.

Inflammations may be local or part of a systemic involvement. Acute pharyngitis may be caused by almost any irritant and typically does not involve an infection by a microorganism. Acute follicular pharyngitis is caused by infectious bacteria, usually streptococci; it is also called septic sore throat. Acute tonsillitis involves the masses of lymphoid tissue found in the back of the pharynx. The tonsils may also be the seat of peritonsillar abscesses. All of the above inflammations may persist as subacute or chronic diseases, but usually the more prolonged forms occur as a result of repeated attacks or low-grade, persistent infections. See STREPTOCOCCUS.

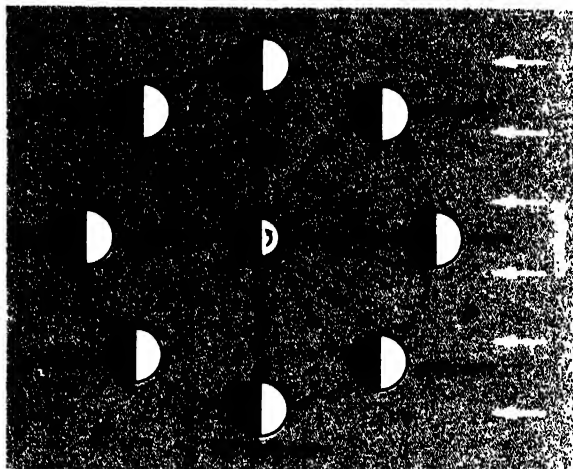
The most common benign tumor of this region is the papilloma; hemangiomas, fibromas, and other less common nonmalignant growths are also found here. The malignant tumors of the pharynx include several varieties of carcinoma and sarcoma, particularly lymphosarcomas in children. See ONCOLOGY.

Nervous disorders seen not infrequently are paresthesias (abnormal sensation) and hyperesthesia (increased sensation). Neuralgia of the glossopharyngeal nerve is marked by severe pain in the neck-ear-jaw region. Motor disorders which affect swallowing may originate from local irritations or from central nervous system disease, as in the case of rabies. See PHARYNX; SOMESTHESIS.

[E.G.ST.]

## Phase (astronomy)

In astronomy, the changing appearance of the Moon, inner planets, and Mars due to the angular difference between the incident light from the Sun and the viewing direction of the observer. Phases of the Moon are a familiar sight. During a lunar month (29.53 Earth days), the Moon completes a cycle of appearances or phases: dark of the Moon, or new Moon, during which the Moon is nearer the Sun than is the Earth; crescent until first quarter (about a week after new Moon); half an illuminated disk, the Moon continues to wax, being gibbous, until it is full; it then wanes through third quarter and completes the cycle as illustrated. A



Viewer from Earth sees Moon differently illuminated as Moon travels around Earth.

solar eclipse by the Moon can occur only at dark of the Moon; a lunar eclipse when the Earth's shadow falls on the Moon can occur only at full Moon.

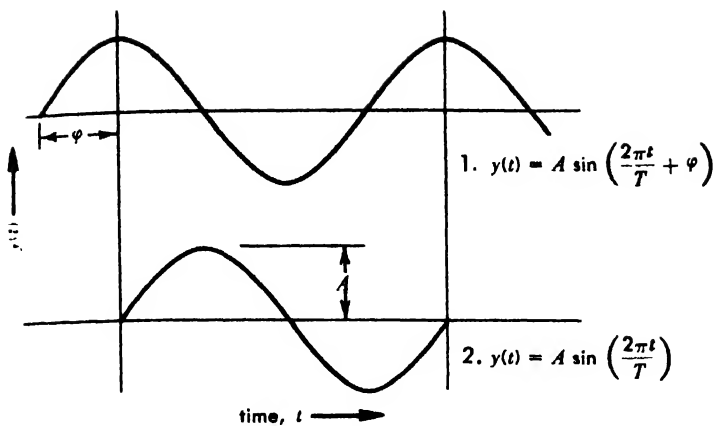
Mercury and Venus show phases like those of the Moon. Mars varies in appearance from full to gibbous. The more remote planets are so far from Earth and Sun that they are viewed from substantially the same angle as that from which they are illuminated and thus show no phase change.

[G.P.K.]

## Phase (periodic phenomena)

The fractional part of a period through which the time variable of a periodic quantity (alternating electric current, vibration) has moved, as measured at any point in time from an arbitrary time origin. In the case of a sinusoidally varying quantity, the time origin is usually assumed to be the last point at which the quantity passed through a zero position from a negative to a positive direction. It is customary to choose the origin so that the fractional part of the period is less than unity.

In comparing the phase relationships at a given instant between two time-varying quantities, the



An illustration of the meaning of phase for a sinusoidal wave. The difference in phase between waves 1 and 2 is  $\varphi$  and is called the phase angle. For each wave,  $A$  is the amplitude and  $T$  is the period.

phase of one is usually assumed to be zero, and the phase of the other is described, with respect to the first, as the fractional part of a period through which the second quantity must vary to achieve a zero of its own (see illustration). In this case, the fractional part of the period is usually expressed in terms of angular measure, with one period being equal to  $360^\circ$  or  $2\pi$  radians. Thus two sine waves of a given frequency are said to be  $90^\circ$ , or  $\pi/2$ , out of phase when the second must be displaced in time, with respect to the first, by  $1/4$  period in order for it to achieve a zero value. See SINE WAVE; see also PHASE-ANGLE MEASUREMENT. [W.J.G.]

## Phase inverter

A circuit having the primary function of changing the phase of a signal by  $180^\circ$ . The phase inverter is most commonly employed as the input stage for a push-pull amplifier. Therefore, the phase inverter must supply two voltages of equal magnitude and  $180^\circ$  phase difference. A variety of circuits are available for the phase inversion. The circuit used in any given case depends upon such factors as the over-all gain of the phase inverter and push-pull amplifier, the possible requirement that the input to the push-pull amplifier may require power, space requirements, and cost. See PUSH-PULL AMPLIFIER.

The over-all fidelity of a phase inverter and push-pull amplifier can be adversely affected by improper design of the phase inverter. The principal problem is that the frequency response of one input channel to the push-pull amplifier may be different from the frequency response of the other channel. The popular phase-inversion circuits are capable of performing the job only after careful selection of components. Some phase-inverter circuits can perform inversion at only one frequency; at other frequencies distortion is introduced because of unequal frequency response characteristics.

**Transformer inverter.** The simplest form of phase-inverter circuit is a transformer with a center-tapped secondary (Fig. 1). Careful design of the transformer assures that the secondary voltages are equal. The transformer forms a good inverter when the inverter must supply power to the grids of the push-pull amplifier. The turns ratio can be adjusted for maximum power transfer. See TRANSFORMER.

The transformer inverter has several disadvantages. It usually costs more and occupies more space than a vacuum-tube circuit. Furthermore, some means must be found to compensate for the frequency response of the transformer, which may not be as uniform as that which can be obtained from vacuum-tube circuits.

**Paraphase amplifiers.** A vacuum-tube amplifier that provides two equal output signals  $180^\circ$  out of phase is called a paraphase amplifier. If coupling capacitors can be omitted, the simplest paraphase amplifier is illustrated in Fig. 2. The same current flows through  $R_L$  and  $R_K$ , and therefore if  $R_L$  and  $R_K$  are equal the ac output voltages from the plate and the cathode must be equal in magnitude and  $180^\circ$  out of phase. The gain of the circuit is less than unity, which is one factor that limits its applicability. A second important factor is that the addition of coupling capacitors and grid-leak resistors, necessary when the circuit is coupled to the push-pull stage, causes the phase inversion to be other than ideal over the frequency range of expected operation.

A paraphase amplifier for which the gain is greater than unity is illustrated in Fig. 3. In the midfrequency range, where the reactance of the coupling capacitors is negligible, the gain of the amplifier section involving  $T_1$  is  $-A_m$  where the minus sign indicates  $180^\circ$  phase shift. If  $R_2/(R_1 + R_2)$  equals  $1/A_m$  the signal at the grid of  $T_2$  will be  $-e_1$ . If the midfrequency gain of the amplifier section involving  $T_2$  is also  $-A_m$ , then  $e_1 = -A_m e_1$  and  $e_2 = A_m e_1$ .

In general,  $e_1 = -A e_1$  and  $e_2 = R_2/(R_1 + R_2) \cdot A^2 e_1$ , where  $A$  is the gain of the stage. This gain is frequency-dependent, and therefore the phase shift varies with frequency. Because  $e_2$  is a function of

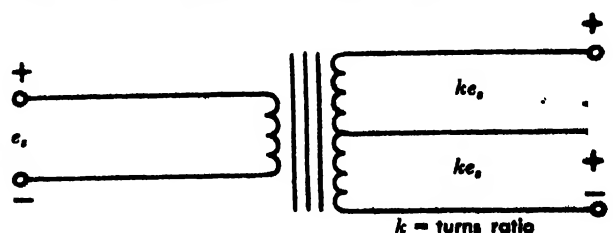


Fig. 1. Transformer as a phase inverter.

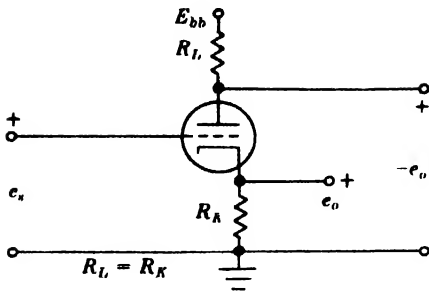


Fig. 2. Single-tube inverter.

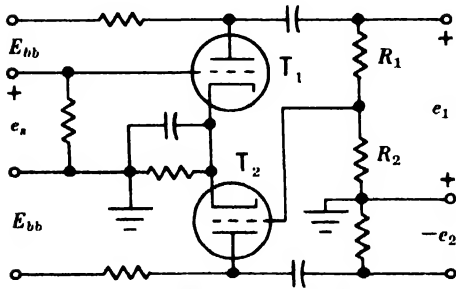


Fig. 3. Two-tube phase inverter.

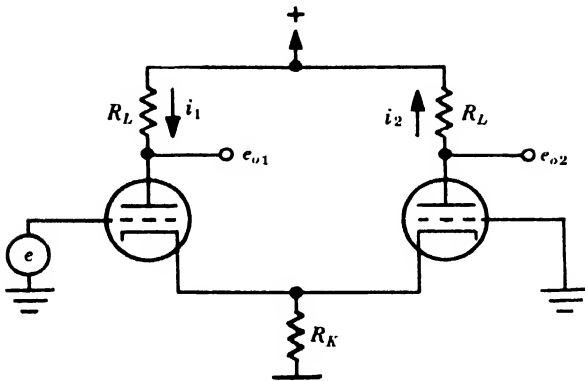


Fig. 4. Cathode-coupled phase inverter.

the square of the gain, the phase shift of  $e_2$  will be twice as large at any frequency as will the phase shift of  $e_1$ . The two voltages are then not  $180^\circ$  out of phase, and distortion will be introduced.

Numerous variations of the above two-tube paraphase amplifier are possible. The two-tube circuit can be quite compact because one twin triode handles the tube requirements. One of the most stable and important paraphase amplifiers is the cathode-coupled phase inverter (Fig. 4). If the cathode resistance  $R_K$  is large compared to the impedance seen looking into the cathode of each tube, the current  $i_1$  will equal  $i_2$ . Under this condition the voltage at one plate is exactly the negative of that at the other plate, and push-pull operation is achieved. The plate-to-plate gain is exactly that which would be provided by a single-tube grounded-cathode amplifier with plate load  $R_L$ .

If the phase-inverter circuit is to produce two voltages  $180^\circ$  out of phase, the equivalent circuits governing the behavior of the two output voltages

must be identical. The midfrequency gain of each must be identical and the phase-shift functions must be identical (which was not true in this circuit). The phase-shift requirements are often compromised in the interests of simplicity of the final circuit and freedom from critical adjustments of key-circuit parameters. [H.F.K.]

## Phase meter

An instrument for the measurement of electrical phase angles. It is sometimes called the crossed-coil meter or the Tuma phase meter and is the basic element of power-factor meters and synchrosopes. When used for power-factor indication it contains two movable coils A and B on a common shaft as shown in the illustration. The two coils move as a unit, the angle  $\beta$  between them remaining fixed. There is no restraining spring acting on the shaft and the system will turn as long as currents produce an average nonzero torque.

The meter contains a fixed coil C which carries the load current. The fixed coil is made in two sections so that its magnetic field will be nearly uniform in the neighborhood of the movable coils.

The instantaneous force induced on coil A is proportional to the product of the instantaneous currents in coils A and C and to the sine of the angle between the planes of A and C. If the currents in A and C are sine-wave currents with phase displacement  $\theta$ , the average torque  $T$  is proportional to

$$(I_a \sin \omega t) [I_c \sin (\omega t + \theta)] \cos \alpha$$

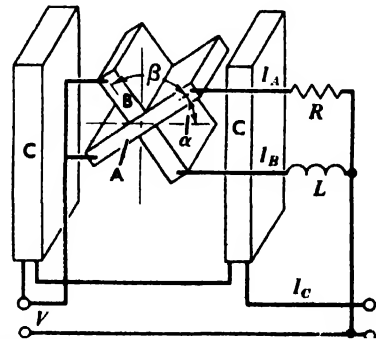
or

$$T = I_a I_c \cos \theta \cos \alpha$$

The movable coil A is placed in a resistive circuit so that its current  $I_a$  is in phase with the voltage  $V$ . The other movable coil B is placed in an inductive circuit so that its current lags the voltage by approximately  $90^\circ$ .

In actual construction the circuit of coil A cannot be made noninductive and the circuit of coil B cannot be made nonresistive. Therefore, the phase angle between the currents in coils A and B is less than  $90^\circ$ . The phase meter is constructed to take this into account.

If the current in coil B leads the current in fixed coil C by the angle  $\theta$ , and the current in coil B lags the current in coil A by the angle  $\phi$ , then the current in coil A will lead the current in the fixed coil by the angle  $(\theta + \phi)$  and the average



Tuma phase meter.

torque on A is proportional to  $\cos(\theta + \phi) \cos \alpha$ , whereas the average torque on B is proportional to  $\cos \theta \cos(\alpha + \beta)$  and in the opposite direction. If these torques do not cancel, the shaft will turn ( $\alpha$  will vary) until the two become equal, or when  $\cos(\theta + \phi) \cos \alpha = \cos \theta \cos(\alpha + \beta)$ .

From this equation it is now evident that if  $\beta$  equals  $\phi$ , then  $\alpha$  equals  $\theta$  and the phase meter indicates directly the phase angle between the sources supplying coil C and the crossed coils. A change of  $5^\circ$  in phase angle produces a deflection of  $5^\circ$  on the phase meter. See PHASE-ANGLE MEASUREMENT; POWER-FACTOR METER. [H.S.O.]

**Bibliography:** F. A. Laws, *Electrical Measurements*, 2d ed., 1938; M. B. Stout, *Basic Electrical Measurements*, 1950.

## Phase modulation

A special kind of angle modulation in which the linearly increasing angle of a sine-wave carrier has added to it a phase angle that is proportional to the instantaneous value of the modulating wave (message to be communicated). Phase modulation (PM) is a scheme for impressing the message to be communicated upon a high-frequency, sine-wave carrier. There is a direct proportionality between the message to be communicated and the phase variations imparted to the modulated wave propagated to the receiver. For basic concepts, technical terms, and supplementary information see MODULATION; see also ANGLE MODULATION.

**Advantages and applications.** Like other forms of angle modulation, PM reduces noise at the cost of extra bandwidth occupancy, transmits constant average signal power, transmits constant peak power which is equal to twice the average signal power, and has a channel-grabbing property whereby if two signals reach the PM detector, the larger signal is accepted to the near exclusion of the smaller.

Important applications include certain types of telegraph, telemetering, and data-processing systems. PM is used in certain microwave radio relay systems, some of which carry telephone conversations and television programs simultaneously. PM techniques are used in many types of measuring and control systems.

The sharp limitation of range (channel grabbing) is especially important for some services. Typical examples using PM include mobile and fixed radio systems for police, airway, and military applications.

**Noise response of PM.** Noise appearing in the output of an angle-modulation detector depends upon the kind of angle modulation, other factors being equal. When the noise disturbance has the characteristics of resistance noise, the average noise power in the output of a PM detector is uniformly distributed with respect to frequency (see NOISE, ELECTRICAL). Under the same conditions, the distribution of root-mean-square noise currents in the output of a frequency-modulation (FM) detector is a distribution increasing linearly with frequency. Other kinds of angle modulation are

characterized by other kinds of noise spectra. Normally, the kind of angle modulation used would be that giving the best signal-to-noise ratio. See FREQUENCY MODULATION.

**Fundamental properties of PM.** Instantaneous phase variations imply and are necessarily accompanied by uniquely related instantaneous frequency variations, and conversely. Also, given one, it is possible to reproduce the other.

For example, in PM, the instantaneous phase variations imparted to the modulated wave are directly proportional to the modulating wave. The resulting variations in instantaneous frequency are, however, directly proportional to the time derivative of the modulating wave.

Similarly, in FM, the instantaneous frequency of the modulated wave is linearly proportional to the modulating wave. However, the resulting variations in instantaneous phase are directly proportional to the time integral of the modulating wave.

Actually, PM and FM are not essentially different. A circuit whose output is inversely proportional to frequency (zero frequency excepted) preceding a phase modulator converts PM to FM, and following an FM detector, converts frequency to phase detection. Similarly, a circuit whose output is directly proportional to frequency (zero frequency excepted) preceding a frequency modulator, converts FM to PM and following a PM detector, converts phase to frequency detection.

Angle modulation manifests itself by the zeros of the angle-modulated wave. These zeros are the exact instants of time that the angle-modulated wave passes through zero. Theoretically, given the zeros it is possible to determine for all values of time, the instantaneous frequency deviations from the fixed frequency of the unmodulated carrier, and also, the instantaneous phase deviations corresponding to the instantaneous frequency deviations. In other words, the zeros, which are nothing more than a distribution of points along the time axis, unambiguously identify the original message.

When detecting an angle-modulated wave perturbed by noise, nonsignificant information must be ignored, because only by this means can the full noise advantage of angle modulation be realized. The limiter in a conventional PM detector ignores nonsignificant information by completely destroying the waveform of the received wave, leaving only the zeros. See PHASE-MODULATION DETECTOR; PHASE MODULATOR. [H.S.BL.]

**Bibliography:** S. Goldman, *Frequency Analysis, Modulation, and Noise*, 1948.

## Phase modulator

An electronic circuit that causes the phase angle of the modulated wave to vary (with respect to the unmodulated carrier) in accordance with the modulation signal. Phase modulators are commonly used in frequency-modulation transmitters in which a phase-modulated wave is altered into a frequency-modulated wave by frequency multiplication. Many methods for the generation of phase-modulated waves are known. Three of the more commonly used

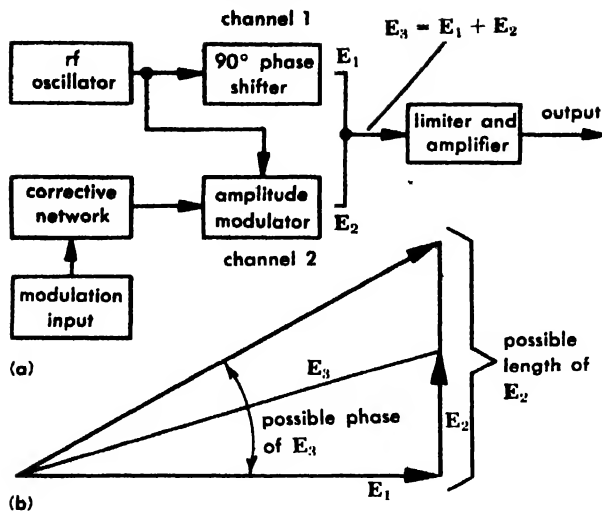


Fig. 1. A method for producing phase modulation. (a) Block diagram showing the principle of operation. (b) Vector diagram showing the combination of voltages. (c) A circuit which permits the necessary functions, except limiting, to be carried out in one tube.

ones are described in this article. See MODULATION; PHASE MODULATION.

**Phase-shift circuit.** The principle of operation of a simple phase-modulator circuit is shown in the block diagram of Fig. 1. Referring to Fig. 1a, the carrier wave is generated by an oscillator, which is crystal-controlled if frequency stability is important. The signal is passed through two channels, 1 and 2. In channel 1 the signal is shifted 90° in phase, and in channel 2 the signal is amplitude modulated. If a frequency-modulated wave is desired, the modulating voltage is first passed through a corrective network in which the output is inversely proportional to frequency. The output of channels 1 and 2 are then combined and passed through the limiter stage to remove the residual amplitude modulation. The action of the circuit is shown in the vector diagram in Fig. 1b. Here the vector  $E_1$  remains constant in amplitude and vector  $E_2$  corresponds to the magnitude of the unmodulated signal. The magnitude of vector  $E_2$  can be changed from zero to twice its normal value as it experiences amplitude modulation. Thus, it can be

seen that the resultant vector  $E_3$  will vary in phase in accordance with the modulating signal. In practice, a phase shift of approximately  $\pm 15^\circ$  can be obtained in this manner, while a nearly constant proportionality is maintained between the modulating voltage and the phase of the output signal.

A simple circuit that permits the generation of the phase-modulated wave in this manner is shown in Fig. 1c. In this particular form, a control-grid-modulated amplifier is used with sufficiently low gain to avoid the oscillations which could otherwise occur as a result of the feedback through the grid-plate capacity. In this case, the action of channel 1 occurs because of the direct transmission of the signal from the rf input to the output through the grid-plate capacity, suffering a 90° phase shift in the process. The output of channel 2 is delivered to the output circuit by the plate current and the normal grid modulation of the amplifier.

**Armstrong circuit.** Another method of obtaining phase-modulated or frequency-modulated waves employs the Armstrong system, shown in Fig. 2, which is a modification of the elementary phase-shift circuit described above. The carrier oscillator supplies two channels. The first undergoes a 90° phase shift as before. The second enters an amplitude-modulation balanced modulator, in the out-

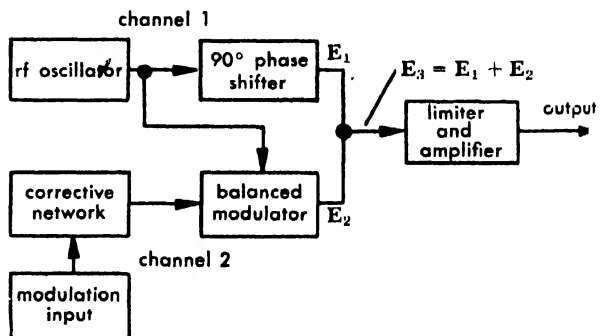


Fig. 2. Block diagram of an Armstrong system of phase (or frequency) modulation.

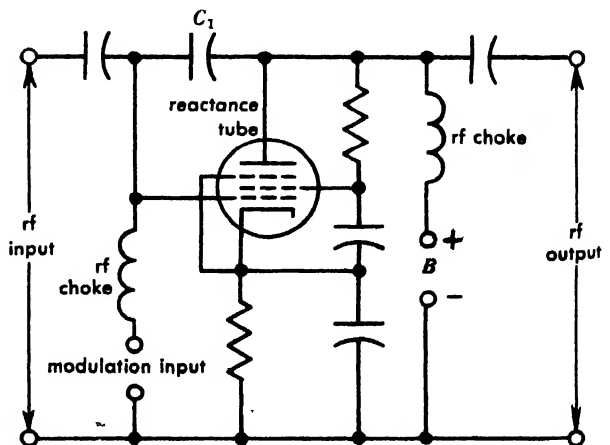


Fig. 3. Circuit diagram of the phase-shifter modulator. The phase shift through the network is varied by controlling the magnitude of the reactance presented by the reactance tube.

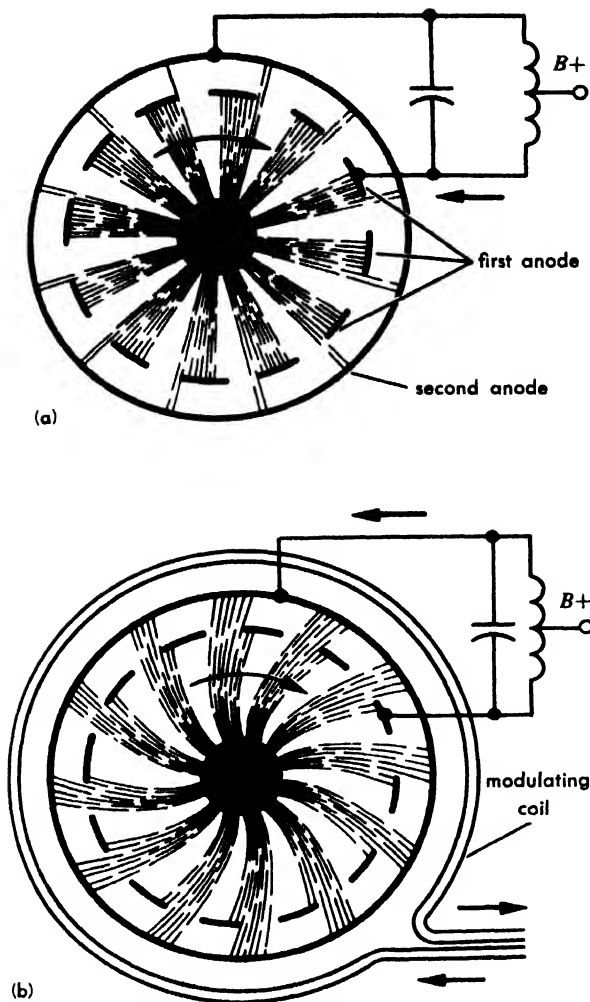


Fig. 4. Diagrams illustrating the principle of operation of the phasitron tube. (a) The rotating electron beams originate from a central cathode (shown in black). (b) Bending of the electron beams by an axial magnetic field, supplied by the modulating coil, advances or retards the arrival of electrons to the second anode. (From R. Adler, *A new system of frequency modulation*, Proc. IRE, 35(1), 1947)

put of which the carrier is suppressed. Upon recombination, the resultant phase-modulated wave is obtained, and has the advantage that the modulation index can be approximately twice as great as in the simple system described previously. A phase shift of  $\pm 30^\circ$  can be obtained with good linearity.

**Phase-shifter modulator.** Another possible way of deriving phase modulation employs a reactance-tube phase shifter in which one of the reactance elements is varied in magnitude electronically with the aid of the modulating voltage. An elementary form of this circuit is shown in Fig. 3. In this circuit, the phase shift is produced by arranging the electron tube to act as a variable reactance, similar to the reactance-tube modulator (see FREQUENCY MODULATOR). The phase shift through the network depends upon the ratio of the reactance of the capacitor  $C$  to the reactance presented from plate-to-cathode of the reactance tube. With a properly designed modulator, it is possible to obtain a nearly linear modulation up to phase shifts of approxi-

mately  $\pm 60^\circ$ . The particular advantage of the phase-shifter type of modulation is that the maximum modulation index can be increased indefinitely by cascading a number of similar phase-shifter modulators.

**Phasitron.** In this system of phase modulation, the phase shift of the carrier is obtained by means of a special electron tube, the phasitron. The principal difficulty with normal phase modulators is a result of the relatively small amount of phase shift that can be produced without introducing nonlinearities and the resultant large amount of frequency multiplication which is usually needed. This difficulty is alleviated to a large degree by means of the phasitron in which a phase swing of as much as  $\pm 200^\circ$  can be obtained with low distortion.

The principle of operation of the phasitron tube can be explained with the aid of Fig. 4. Suppose that an electron stream is produced in the form of wheel spokes from a central cathode in which the electrons move outward along radial lines and rotate with uniform velocity. Figure 4a shows that the electron beams alternatively will be intercepted by the first anode or permitted to pass to the second anode. The tuned circuit connected between the two anodes is therefore excited at a constant frequency equal to the speed of rotation of the electron spokes by the segmented first anode. If a magnetic field is applied parallel to the axis of the cathode, the electron beams are deflected as shown in Fig. 4b and arrive at the second anode either earlier or later than normal, depending upon the direction of the magnetic field. The advance caused by the magnetic field corresponds to the width of one segment of the first anode, so that the second-anode current is advanced in phase by  $180^\circ$ . The amount of phase shift that can be obtained depends upon the magnetic field produced by the modulating signal.

In the phasitron, the rotating electron spokes are produced by surrounding the cathode by a squirrel cage of conductors to which three-phase voltages derived from a crystal oscillator are applied. A possible arrangement of such wires is shown in Fig. 5. If all the wires numbered 1 are connected together, and all number 2, and all 3, and the three sets are excited from a three-phase voltage source, a ro-

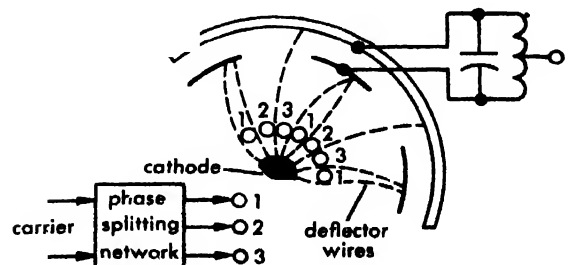


Fig. 5. Arrangement of deflector wires in the vicinity of the cathode of the phasitron which permits generation of rotating electron beams. Three-phase potentials are applied to the three sets of wires. (From R. Adler, *A new system of frequency modulation*, Proc. IRE, 35(1), 1947)



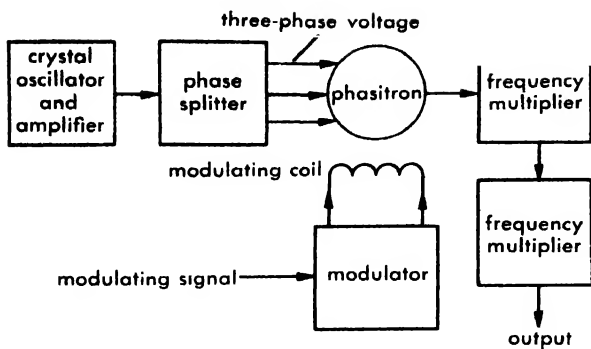


Fig. 6. Block diagram indicating the method of producing frequency modulation by employing the phasitron tube.

tating electron stream can be produced. This can be seen by imagining that a set of wires numbered 1 are positive whereas 2 and 3 are negative; this causes most of the electron current to be emitted in the direction of the wire numbered 1, thus causing the spoke to form. The field along the squirrel cage rotates in a manner similar to that in a synchronous motor.

Figure 6 shows the block diagram of a frequency-modulated transmitter using a phasitron tube. A crystal oscillator operating at some low radio frequency supplies a voltage to the network consisting of inductances and capacitors from which three voltages  $120^\circ$  apart are derived. The introduction of this three-phase voltage to the squirrel-cage elements of the phasitron creates the rotating electron beam. The rf output from the phasitron is introduced into subsequent frequency-multiplier stages. The phase of the output carrier is shifted by means of the signal derived from a modulator as indicated. [E.L.G.]

**Bibliography:** R. Adler, A new system of frequency modulation, *Proc. IRE*, 35(1):25-31, 1947; F. M. Bailey and H. P. Thomas, Phasitron F-M transmitter, *Electronics*, 19(10):108-112, 1946; H. S. Black, *Modulation Theory*, 1953.

## Phase velocity

The velocity of propagation of a simple harmonic wave. Such a wave is propagated with constant waveform, and the phase velocity is the speed, measured in a direction normal to the wavefront, at which one must move to keep up with a place of the same phase. The phase velocity  $v_p$  is given by  $v_p = \lambda f$ , where  $\lambda$  is wavelength and  $f$  frequency.

The magnitude of phase velocity is determined by the intrinsic properties of the medium in which the wave is propagated and by the actual mode of propagation. For small-amplitude acoustical waves in a quiescent, unbounded, gaseous medium, the phase velocity is essentially independent of frequency and equal to the speed of sound  $c$  in the medium. This speed is given by the equation  $c^2 = \gamma P_0 / \rho_0$ , where  $\gamma$  is the ratio of specific heat at constant pressure to that at constant volume,  $P_0$  is the ambient pressure in the medium, and  $\rho_0$  is its ambient density. In a perfect gas, the speed of

sound becomes a function only of the absolute temperature of the medium. In air, this quantity is approximately  $c = 49.03\sqrt{R}$  ft/sec, where  $R$  is in degrees Rankine, or  $c = 20.05\sqrt{T}$  m/sec, where  $T$  is in degrees Kelvin. The phase velocity of electromagnetic waves also depends on the medium, but in a vacuum, the velocity  $c$  is a universal constant approximately equal to  $3 \times 10^8$  m/sec.

Waves in a medium which require higher than second-order equations for their description, such as lateral vibrations of a bar, generally have phase velocities which are dependent upon frequency. Thus, waves of different frequencies travel with different velocities, resulting in dispersion. See GROUP VELOCITY; WAVE MOTION. [W.J.G.]

## Phase-angle measurement

The determination of the relative times at which alternating currents and voltages in a circuit take on zero values. If two voltages  $v_1$  and  $v_2$  are zero at the same instant, they are in phase, with zero phase difference (or out of phase with  $180^\circ$  difference). If one voltage  $v_1$  passes through zero  $\frac{1}{8}$  cycle before a second voltage  $v_2$ , it leads by  $360^\circ/8$  or  $45^\circ$  (see Fig. 1). The common phase meter, a commercial device for determining the angle between current and voltage, can be used when its presence will not disturb the circuits under measurement (see PHASE METER). When the phase angles to be measured are in high-impedance or low-power circuits, this instrument is unsatisfactory and other measurement methods must be employed.

**Three-voltmeter method.** This method can be used when the voltages involve a common point. Figure 2a shows three terminals  $a$ ,  $b$ , and  $c$ . If the voltages  $v_{ab}$ ,  $v_{bc}$ , and  $v_{ca}$  are measured by a high-impedance voltmeter (one voltmeter is sufficient) the magnitudes can be plotted to give a triangle, Fig. 2b. The angle  $\theta$  between  $v_{ab}$  and  $v_{bc}$  can be determined from the law of cosines in trigonometry,

$$v_{ca}^2 = v_{ab}^2 + v_{bc}^2 + 2v_{ab}v_{bc} \cos \theta$$

**Electronic phase-angle meter.** This instrument gives the angle  $(\pi - \theta)$  of Fig. 2 directly. One

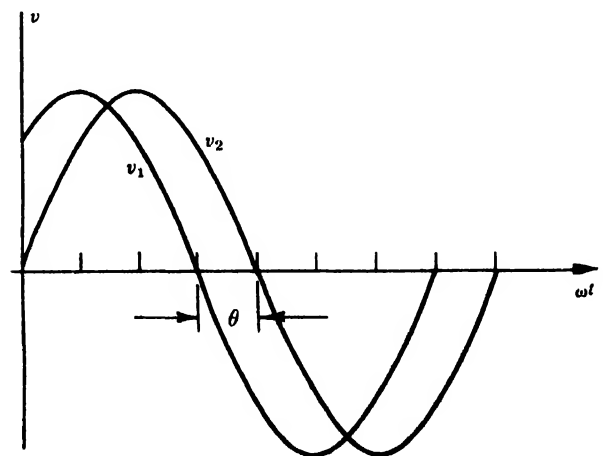


Fig. 1. Phase angle between two voltages.

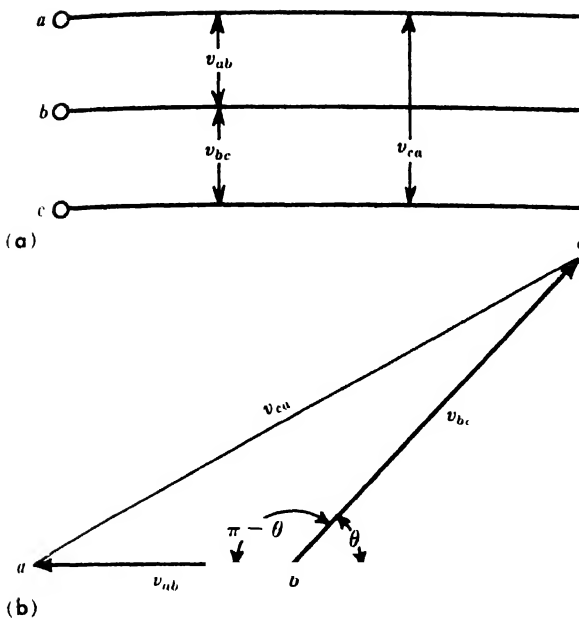


Fig. 2. Voltages employed in three-voltmeter method. (a) Circuit diagram. (b) Vector diagram.

such instrument converts the two voltage waves to square waves by repeated amplification and limiting. The zero crossings of the square waves are identical to the zero crossings of the original voltage waves. The two square waves are applied to the input of a circuit that will pass current only when both square waves are positive. In this case the greater the lag of one voltage, the smaller the overlap of the positive portions and the lower the average current. The current in this case is proportional to  $\theta$  of Fig. 2.

This circuit has the theoretical limitation that each input voltage must be greater than a critical minimum value. In practice the critical value is determined by the noise on the amplifier input. If the voltage is too low, this noise causes a random zero-crossing shift and the results would be subject to this uncertainty.

A precision phase-angle meter for high-frequency voltages uses a variable delay line, and its operation is based on the fact that the difference of two voltages of constant amplitude is a minimum when the two are in phase. One of the two voltages to be compared is connected to both inputs of a variable-delay line, which is then adjusted to give a minimum output. The two voltages to be compared are then connected to the two terminals and the delay line is readjusted to give a minimum output. The change in the delay-line setting gives the time delay of one voltage relative to the other. When the frequency is known, the time delay can be computed as angle of lag. If  $\Delta t$  is the change in the delay-line setting and the frequency is  $f$  cycles per second, the phase angle is given by  $2\pi f \Delta t$  radians or  $360 f \Delta t$  degrees.

**Oscilloscope methods.** Phase-angle measurements by oscilloscopes are popular in the laboratory when quick approximate results are required. If

one voltage is connected to the vertical amplifier and the other to the horizontal amplifier, a Lissajous figure is obtained.

If the two voltages are of the same frequency, as is the case when phase angles are measured, the basic figure is an ellipse. A straight line with a positive slope implies that the two waves are in phase. If one leads the other, the cathode-ray beam starts back in one direction before it reaches a maximum in the other and the ray traces an ellipse.

If the amplifiers are adjusted so that the horizontal amplitude is equal to the vertical amplitude, the slope of the straight line for in-phase signals is  $45^\circ$ . The ellipse widens to a circle when the phase angle is  $90^\circ$ . Intermediate values are indicated by the formula

$$\tan \frac{\phi}{2} = \frac{b}{a}$$

where  $\phi$  is the phase angle sought,  $b$  is the width of the ellipse, and  $a$  is its length. See LISSAJOUS FIGURES: OSCILLOSCOPE, CATHODE-RAY.

Another method of utilizing an oscilloscope, developed by the author, may be illustrated by the following considerations. If a signal is applied across

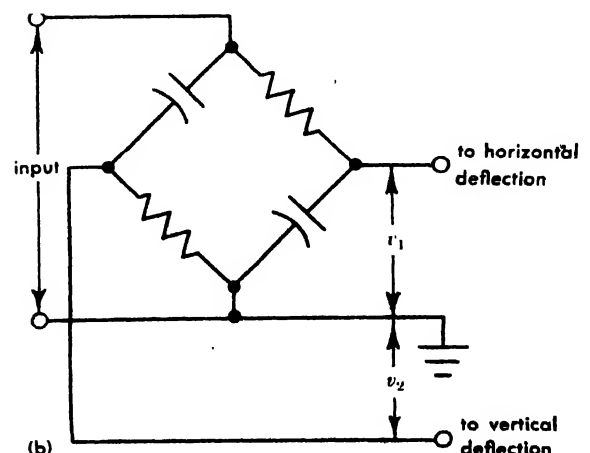
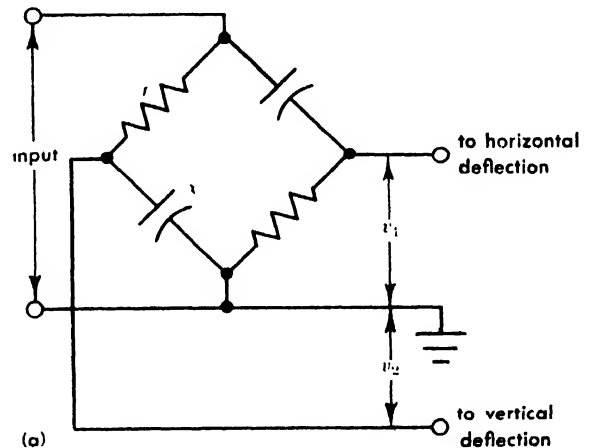


Fig. 3. Basic circuits for measuring phase angle with an oscilloscope. (a) Clockwise circular sweep generated. (b) Counterclockwise circular sweep generated.

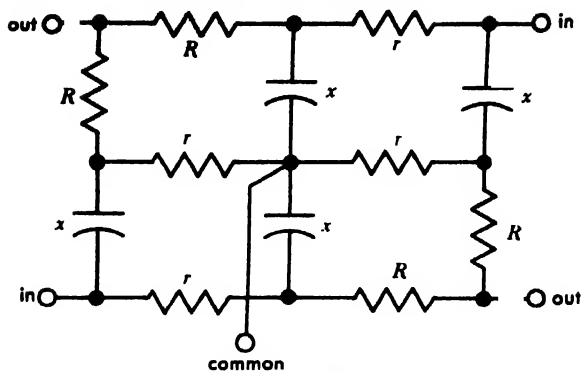


Fig. 4. Actual circuit for measuring phase angle with an oscilloscope.

the input of the bridge circuit shown in Fig. 3a, the output voltages  $v_1$  and  $v_2$  will be  $90^\circ$  apart, and if  $r$  equals  $x$  (equals  $1/2\pi fC$ ) they will be equal in amplitude. Now if  $v_1$  and  $v_2$  are applied to the vertical and horizontal amplifiers of an oscilloscope, the resulting trace will be a circle. If  $r$  and  $x$  are interchanged as in Fig. 3b the trace will again be a circle, but the spot will trace the circle in the opposite direction.

If these two circuits are combined so that the sums of the outputs are applied, the spot cannot go around the circles in opposite directions at the same time, and it will trace a straight line instead. As the phase of one voltage is advanced, say by the angle  $\Delta\theta$ , the straight line will rotate on the screen through the angle  $2\Delta\theta$ . A scale can be marked on the screen and either end of the straight line used for reference. The presence of harmonics and slight errors in the  $r = x$  relationship or in the equality of the input voltages will cause the line to open into a narrow ellipse. The slope of the major axis of the ellipse in this case is used for the slope of the straight line. Figure 4 shows a working circuit. It is built so that  $r = x$  at the operating frequency, and  $R$  is several times  $r$  so that the voltages are added without appreciable loading of the bridge circuits.

**Electronic switch.** An electronic switch can be used with an oscilloscope as a phase-angle meter. The switch permits the oscilloscope to display first one wave and then the other. If the linear sweep is at the same frequency as the waves being compared, the two waves will be superimposed on the screen. The phase difference and the period can be measured on the screen in inches, the ratio giving the phase angle as a fraction of  $360^\circ$ .

**Phase-order indicators.** These devices are used to indicate which phase voltage of a polyphase circuit leads or lags another. If the voltage vectors of a three-phase generator are as indicated in Fig. 5a, the voltage from neutral to line 2 reaches a maximum  $1/3$  cycle (or period) after the voltage of line 1 and  $1/3$  cycle before the voltage of line 3. The voltage of phase 2 lags that of phase 1 and leads that of phase 3. The phase order or phase sequence is then said to be 1-2-3.

Relative motion between the armature conductors and the magnetic field induces the voltages in an alternator. Therefore, if the alternator were rotated in the opposite direction, the order in which the phase voltages reached maximum would be reversed. The phase sequence would be 1-3-2 as shown in Fig. 5b.

A miniature three-phase motor designed to rotate clockwise when connected to a three-phase system possessing a phase sequence 1-2-3 is used as a phase-sequence indicator. Counterclockwise rotation would indicate a 1-3-2 phase sequence.

A common type of phase-order indicator consists of an inductance and two lamps connected in Y to the three-phase line as in Fig. 6a. If we assume that the inductive reactance is very high, the common connection on the Y is at a voltage nearly equal to the midpoint of line 2-3 in Fig. 6b and c. The voltage across the reactor is  $v_{n1}$ , and the current lags by  $90^\circ$  and lies on  $n2$  in Fig. 6b and on  $n3$  in Fig. 6c. This current divides, part going through each lamp. The result is to increase the current in lamp 2 when the phase sequence is 1-2-3, and to increase the current in lamp 3 when the phase sequence is 1-3-2.

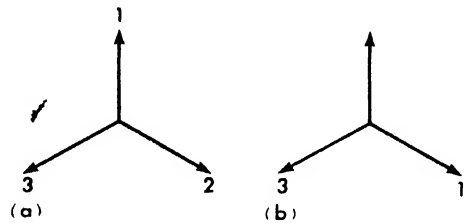


Fig. 5. Phase sequence. (a) Sequence 1-2-3. (b) Sequence 1-3-2.

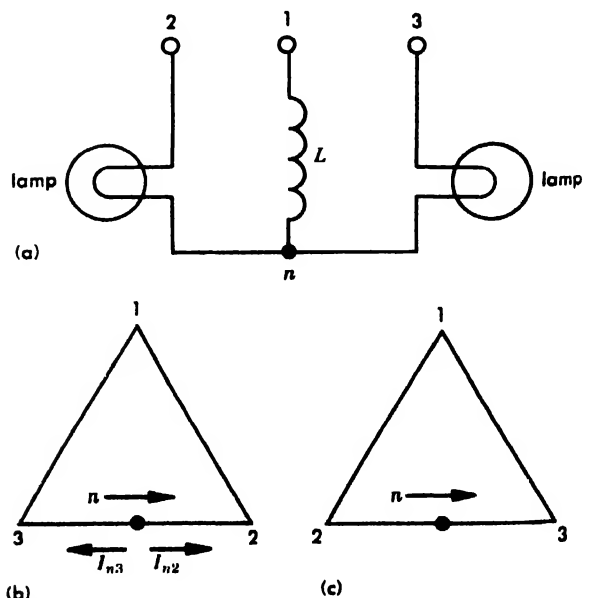


Fig. 6. Phase-sequence indicator. (a) Circuit diagram. (b) Vector diagram for sequence 1-2-3. (c) Vector diagram for sequence 1-3-2.

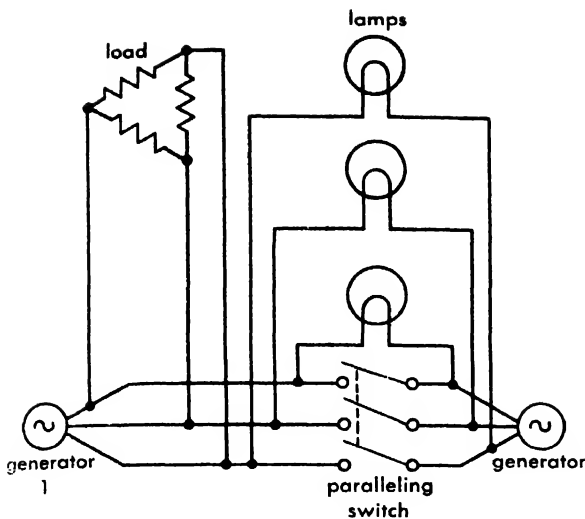


Fig. 7. Paralleling a second generator with a first already under load.

**Phase-relation indicators.** These devices are used to indicate the instant when two generators or sources of alternating voltage are in phase with one another. If two voltages reach maximum at the same time, they are in phase. When two sources are to be connected in parallel they should have the same voltage, frequency, and phase. A voltmeter and a tachometer can be used to indicate when the voltages and frequencies are nearly equal. The phase relation between the two sources is shown by means of phasing lamps or by means of a synchroscope or synchronizer.

Phasing lamps placed across the open switch used to parallel two generators will often suffice to indicate an in-phase condition (see Fig. 7). Depending upon the relative phase of the two machines, the lamp voltage varies from the sum to the difference of the machine voltages. As the two frequencies approach one another, the lamps flicker, changing from full bright to dim at a decreasing rate. If the two frequencies are equal, the lamp will maintain a fixed brilliance. Usually the oncoming machine is set with a slightly higher frequency so that it will take up some load rather than be an additional load on the system. As the lamps slowly go through the dim phase, the switch is closed, connecting the machine to the system.

**Synchroscope or synchroscope.** This is a variation of the Tuma phase meter. The current in the fixed coil is supplied by one machine; the current in the movable crossed coils is supplied by the other machine. If the two machines are in synchronism, their frequencies are equal and the crossed coils will take a position depending upon the relative phase angle. If the frequency of one machine is slightly higher, the phase will continue to vary and the crossed coils will rotate in a direction determined by whether the speed is too low or too high. Generally the incoming machine is given a slightly higher speed and is connected to the line when the synchroscope pointer drifts past the zero

mark. See ELECTRICAL MEASUREMENTS; POWER-FACTOR METER; WATTMETER. [H.S.O.]

**Bibliography:** G. R. Partridge, *Principles of Electronic Instruments*, 1958; M. B. Stout, *Basic Electrical Measurements*, 1956.

### Phase-modulation detector

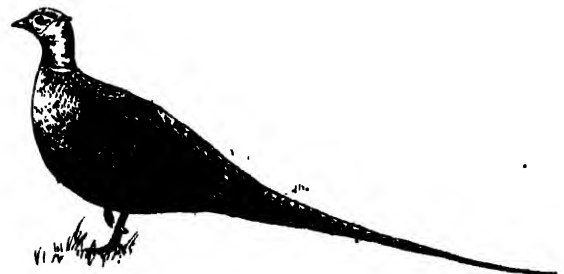
A device for the detection of phase-modulated radio signals. Phase-modulated signals can be detected in a manner identical to that used with frequency-modulated signals because there is no essential difference between these forms of modulation. However, because the modulation index varies differently with modulation frequency, frequency-modulation detectors need to be modified by the addition of a low-pass network in which the amplitude is made to vary inversely with the modulation frequency. With this addition, any frequency-modulation detector can be made to operate satisfactorily for the demodulation of phase-modulated waves. See FREQUENCY-MODULATION DETECTOR.

To demodulate a phase-modulated wave without amplitude distortion, the signal derived from a frequency-modulation detector must be corrected by passing the output through a low-pass filter in which the output is inversely proportional to the modulation frequency. If, for example, it is desired to handle a bandwidth of 50–20,000 cycles per second, a 400 to 1 difference in transmission must be supplied by the corrective network. This large difference in signal strength is difficult to obtain from all normal detector circuits without frequency distortion and other noise.

In certain phase-modulation communication systems the difficulty just mentioned is partially eliminated by employing transmitter preemphasis, that is, decreasing the degree of modulation at the higher modulation frequencies. If used, the inverse of the preemphasis networks must be employed at the detector in addition to the corrective network mentioned. See DETECTOR. [E.L.G.]

### Pheasant

Any of many fowl-like birds of the family Phasianidae, primarily occurring in southeastern Asia, with numerous, beautiful species native to southeastern China, Tibet, Burma, and the Malay Peninsula. Pheasants are characterized by striking plumage, long tails, spurred legs, and the absence



The ring-necked pheasant, *Phasianus colchicus*; length to 3 ft. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

of feathers on the sides of the head. They eat grain, weed seeds, berries, insects, and snails. Pheasants have been raised by man for many centuries, first reaching western Europe through introduction by the Romans.

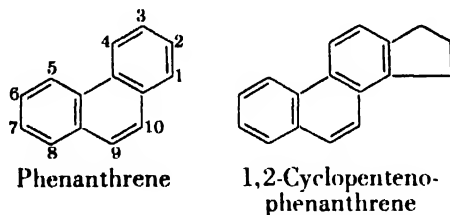
The ring-necked pheasant, *Phasianus colchicus*, has been repeatedly introduced into the United States and has become well established over a wide area of the northern United States and southern Canada, where it ranks as a major game bird. See GALLIFORMES. [J.D.B.]

## Phenacetin

One of a general class of medicinals known variously as analgesics, antifebrins, or antipyretics, of which acetanilide is the best known. Phenacetin is an acetyl derivative of *p*-phenetidine. It is made by the reaction of  $\text{NaO}-\text{C}_6\text{H}_4-\text{NO}_2$  with  $\text{C}_2\text{H}_5\text{Cl}$  to form  $\text{C}_2\text{H}_5\text{O}-\text{C}_6\text{H}_4-\text{NO}_2$ , which is reduced to the corresponding amine and acetylated to form phenacetin,  $\text{C}_2\text{H}_5\text{O}-\text{C}_6\text{H}_4-\text{NHCOCH}_3$ . The reaction is continued as a cyclic process in which phenol, acetic acid, and ethyl chloride are continuously supplied. Phenacetin is less toxic than acetanilide, but it lowers the ability of blood to combine with oxygen. See ASPIRIN; HYPOTHERMIA. [A.L.H.]

## Phenanthrene

A colorless, crystalline hydrocarbon ( $\text{C}_{14}\text{H}_{10}$ ) which melts at about  $100^\circ\text{C}$  and boils at  $332^\circ\text{C}$ . Phenanthrene is usually obtained from coal tar,



but it may also be produced by the hydrogenation of coal. Since carbazole and anthracene (usually present in crude phenanthrene) form mixed crystals with it, commercial grades of phenanthrene usually melt at higher temperatures than the pure compound.

Phenanthrene may be hydrogenated in the presence of copper chromite to yield 9,10-dihydrophenanthrene, or it may be oxidized to yield 9,10-phenanthrenequinone. In general, substitution reactions yield a mixture of products which are difficult to separate.

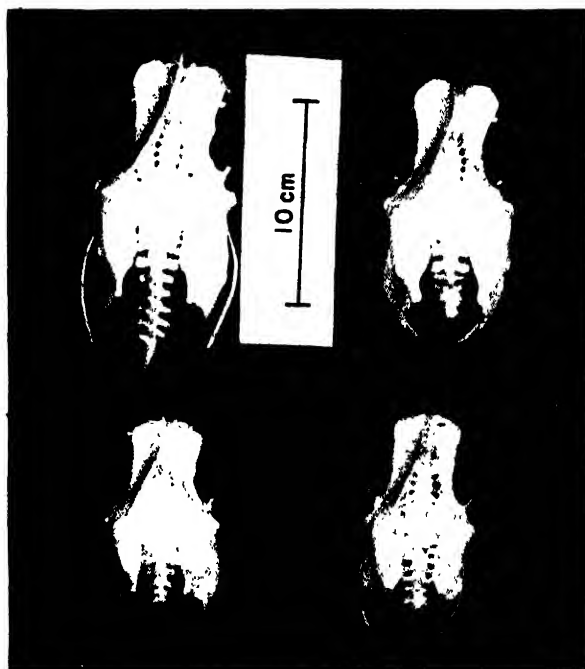
Phenanthrene has little commercial importance. It is of interest because the nucleus is found in some resin acids and is produced by the degradation of certain alkaloids. Reduction products of 1,2-cyclopentenophenanthrene may be regarded as forming the skeleton of the steroids. See AROMATIC HYDROCARBON; POLYNUCLEAR HYDROCARBON; STEROL. [C.K.B.]

## Phenocopies

Nonhereditary variations of form or function resembling mutant traits, but caused by external conditions during development. They occur presumably in all types of organisms, spontaneously or after experimental intervention. Present knowledge of phenocopies is derived chiefly from work on *Drosophila melanogaster* and chicken embryos.

In hereditary (dominant or recessive) rumplessness in the chicken as well as in the insulin-produced phenocopies, the abnormalities vary from complete absence of all tail vertebrae to conditions deviating only slightly from the normal one. The specimens shown in the figure are of different ages and sizes but illustrate an intermediate condition with abnormal and fused tail vertebrae. These intermediate conditions are caused by the presence of modifying genes which prevent the mutant genes or the phenocopy-inducing agent from exerting their full force. Developmental studies have shown that the insulin-induced phenocopy is more closely akin to recessive than to dominant rumplessness.

The phenocopy-producing agents may be physical, for example, heat shocks, x-rays, or chemical agents. The principal response-determining factors are the developmental stage at exposure, the force of external agent, and the genetic constitution of responding organism. Unrelated phenocopies may be produced in different stages by the same agent and by different agents in the same stage. The force or dosage of the external agent is chiefly a source



Pelvic bones and tail vertebrae of a normal chicken (upper left), one with dominant rumplessness (upper right), one with recessive rumplessness (lower left), and one produced by injecting insulin into the developing egg after 4 days of incubation (lower right).

of quantitative variations in the incidence and type of response, but qualitative response differences may occur. The genetic constitution of the exposed organism is often decisive in determining the effect. Plus or minus selection is possible. The presence or absence and the kind of modifying genes (residual heredity) play an important role, especially the presence of genetic factors predisposing to specific responses. Organisms can often be protected against phenocopy-inducing chemical agents by specific supplements. Weaknesses of genetic integration (homeostasis) and interference in enzyme systems appear to be major factors. See GENE ACTION. [W.L.]

**Bibliography:** W. Landauer, On phenocopies, their developmental physiology and genetic meaning. *Am. Naturalist*, 92(865):201-213, 1958.

## Phenocryst

A relatively large crystal embedded in a finer-grained or glassy igneous rock. The presence of phenocrysts gives the rock a porphyritic texture. Phenocrysts are represented most commonly by feldspar, quartz, biotite, hornblende, pyroxene, and olivine. Strictly speaking, phenocrysts crystallize from molten rock material (lava or magma). They



Granite (quartz monzonite) from the Sierra Nevada of California showing numerous phenocrysts of microcline feldspar in parallel orientation with banded structure of the rock. The phenocrysts appear to have replaced the rock and are called porphyroblasts. Hammer is 10 in. long. (USGS photograph by W. B. Hamilton)

commonly represent an earlier and slower stage of crystallization than does the matrix in which they are embedded. Phenocrysts are to be distinguished from certain relatively large crystals (porphyroblasts) which develop late in solid rock as the result of metamorphism or metasomatism. If the origin of a large crystal is in question, the non-genetic term megacryst should be used. See AUREOLE; CONTACT; IGNEOUS ROCKS; PORPHYROBLAST; PORPHYRY; RAPA-KIVI GRANITES. [C.A.CA.]

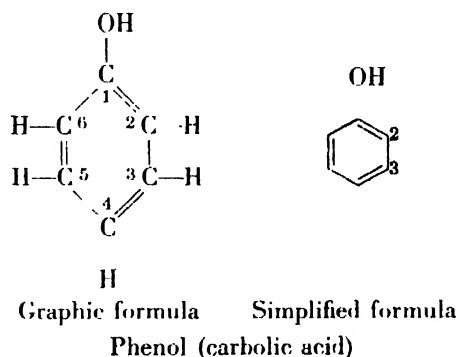
## Phenol

One of a class of acid organic compounds whose common structural feature is a hydroxyl group attached directly to an aromatic-ring system.

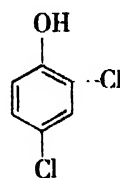
Phenol (hydroxybenzene) is toxic to all types of living cells. In concentrated form, it produces severe skin burns. Phenols in general are toxic to microorganisms. Phenol itself, in dilute solution, was the first deliberately used antiseptic, and many phenols are now used for this purpose. Picric acid is used to treat burns.

**Uses.** The chief use for phenol is in the manufacture of phenol-formaldehyde resins and plastics, but it is also a raw material required for the preparation of more complex phenols, which are used as drugs, disinfectants, antiseptics, fungicides, insecticides, wood preservatives, antioxidant preservatives for gasoline, oils, rubbers, fats, and foodstuffs, and in the manufacture of dyes, plasticizers, weed-killers, detergents, and epoxy resins. The annual production of phenol in the United States now exceeds 500,000,000 lb.

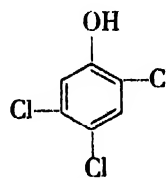
**Structure and nomenclature.** The simplest member of this class is phenol, also known as carbolic acid. Structures of other members of the class are derived from that of phenol by replacing one or more of the hydrogen atoms attached to carbon atoms with other atoms or atomic groupings. The classification of phenols is based on the number and nature of these groupings (other than hydroxyl) that are attached to the 6-membered ring of carbon atoms. The names of phenols are based



on the assignment of numbers to the carbon atoms of the aromatic ring. The carbon atom attached to the hydroxyl group is assigned the number 1, and the five remaining carbon atoms are then numbered consecutively. If but one atom or group is located on the aromatic ring at position 2, this group is said to be ortho (*o*) to the hydroxyl group; if the second group is located at position 3, it is meta (*m*); at position 4, it is para (*p*). If two or more atoms or groups in addition to the hydroxyl are present, their positions are designated by the appropriate numbers.

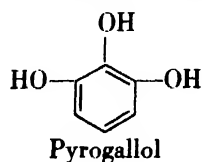


2,4-Dichlorophenol



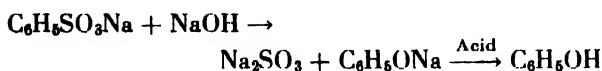
2,4,5-Trichlorophenol

Phenols whose structures contain more than one hydroxyl group bound to an aromatic ring system are called polyhydric phenols. Examples are catechol, resorcinol, and hydroquinone, all dihydric phenols. Pyrogallol, a trihydric phenol, is a photographic developer. Phenols in which one or more

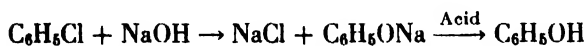


hydroxyl groups are joined directly to an aromatic system comprising two rings of carbon atoms are represented by the naphthols. See NAPHTHOL.

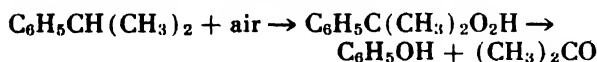
**Production.** Phenol is produced by four competing processes: (1) The benzenesulfonate process involves the sulfonation of benzene by concentrated sulfuric acid and then fusion of the sodium benzenesulfonate with caustic soda:



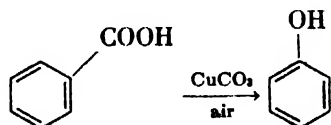
(2) In the chlorobenzene (Dow) process chlorobenzene is treated with caustic soda in the presence of copper at elevated temperatures and pressures:



(3) In the Raschig process benzene is first converted to chlorobenzene by hydrochloric acid and air in the presence of a catalyst, and then the phenol is formed from chlorobenzene and water in the presence of another catalyst. (4) In the cumene peroxidation process cumene hydroperoxide, formed by the action of air on cumene, is converted by sulfuric acid into phenol and acetone:

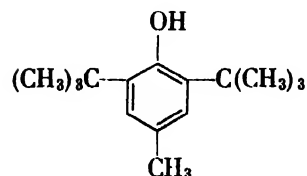
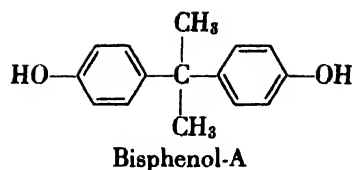


A more recent process consists of heating benzoic acid in air in the presence of copper carbonate.

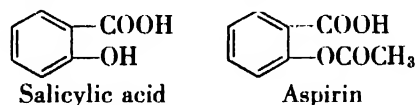


Less than 5% of the phenol now manufactured in this country is obtained directly from coal tar. More complex phenols are prepared from phenol itself or by the adaptation of one of the methods used to prepare phenol to appropriately constituted aromatic compounds.

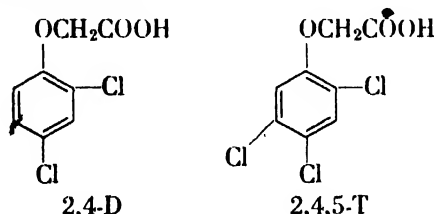
**Complex phenols.** Bisphenol-A is made from phenol and acetone and is an intermediate in the preparation of epoxy resins. 2,6-Di-*tert*-butyl-4-methylphenol is made from *p*-cresol and isobutylene, and is used to protect gasoline, oils, rubber, and foods from deterioration caused by atmospheric oxygen. Pentachlorophenol is used to protect wood against attack by fungi and termites. Its sodium salt is used to treat industrial water to pre-



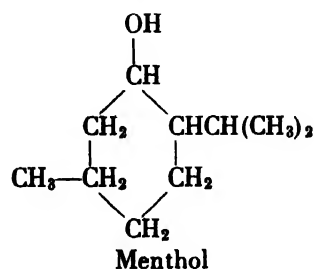
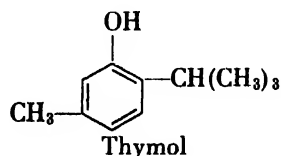
vent the growth of slime and algae. Salicylic acid is an intermediate from which aspirin and other analgesic drugs are made. 2,4-Dichlorophenol and



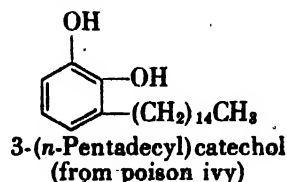
2,4,5-trichlorophenol are intermediates in the manufacture of the weed-killers 2,4-dichlorophenoxyacetic acid (2,4-D) and 2,4,5-trichlorophenoxyacetic acid (2,4,5-T) which are usually used as their salts or esters.



Some phenols occur in natural products. For example, thymol, from thyme oil, is used in dilute solution as an antiseptic and as an intermediate from which menthol is prepared. The toxic agents



in poison ivy are related to catechol; one of them





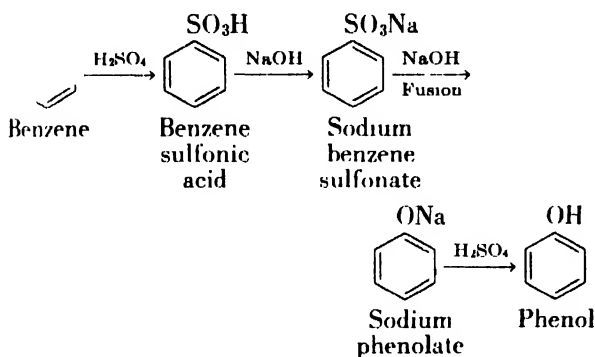
is 3-(*n*-pentadecyl)catechol. See PARA-AMINOPHENOL; CATECHOL; CRESOL; HYDROQUINONE; PICRIC ACID; RESORCINOL.

[R.B.C.]

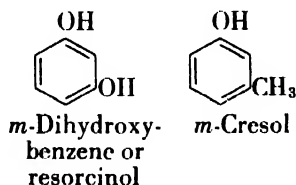
## Phenol-formaldehyde resin

One of the condensation products of phenols or phenolic derivatives with aldehydes such as formaldehyde and furfural. The term phenoplasts is sometimes used to refer to the whole group of products. The phenol-formaldehyde resins, developed commercially between 1905 and 1910, were the first truly synthetic polymers and have found wide usage for electrical insulation, molded objects, shell molds for metals, laminates, adhesives, and many other applications. They are characterized by low cost, high strength, and resistance to aging.

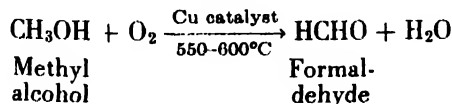
Phenol is prepared by the direct oxidation of benzene, by the hydrolysis of chlorobenzene, or by the alkali fusion of sodium benzene sulfonate:



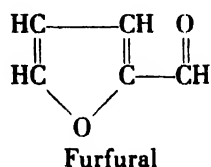
Resorcinol, obtained by the alkaline fusion of *m*-benzene disulfonic acid, and *m*-cresol from coal tar are also used.



Formaldehyde is produced by the oxidation of methane or methyl alcohol.

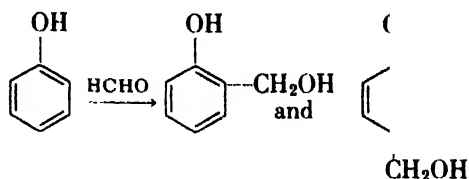


and furfural is obtained by the hydrolysis of oat hulls.



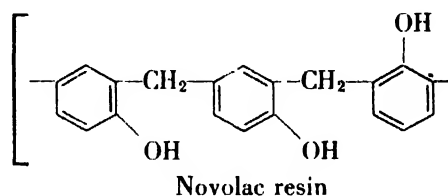
**Polymerization.** In the presence of alkali, phenol and aqueous formaldehyde react to form a solution of phenolic alcohols or methylol derivatives

with the methylol groups in the ortho and para positions.



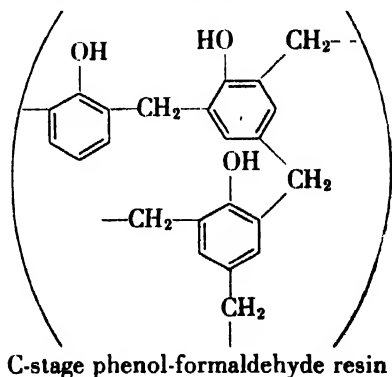
The methylol phenols are soluble, fusible materials which are generally called the A-stage resin.

In the presence of acid and less than 0.86 moles of formaldehyde per mole of phenol, the primary alcohols react to yield diphenylmethane polymers called novolacs which are soluble and fusible and contain 4-20 phenol units.



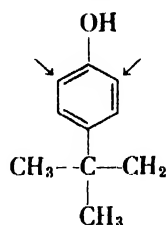
By further heating of the A-stage resin or by the addition of more formaldehyde to the novolac, further condensation takes place with the formation of the B-stage product, a brittle resin that is partially soluble and fusible.

In the production of phenolic-resin molding compositions, it is common practice to neutralize, concentrate, and dry the B-stage resin, to mix it with fillers and curing agent, and then to compact it into the form of pellets or briquets. The curing agent is hexamethylenetetramine, which at the temperature of molding reacts with water to form formaldehyde and ammonia. In the presence of ammonia and the additional formaldehyde, the B-stage resin cures in the mold to yield a highly cross-linked, insoluble, and infusible C-stage product.



By use of *m*-phenol derivatives such as resorcinol or *m*-cresol, resins are obtained which cure rapidly at low temperature because the meta substituents activate the ortho and para positions. By use of an ortho or para alkyl phenol, such as *p*-tert-butylphenol, which has only two active sites avail-

able for reaction (difunctional),



oil-soluble, thermoplastic resins are formed instead of the cross-linked materials obtained from the tri-functional phenols just discussed. These products, being somewhat more expensive than the ordinary phenolic resin, are used in special paint, varnish, and adhesive formulations.

With the use of furfural instead of formaldehyde, the B-stage resin has the unique property of remaining thermoplastic for a relatively long time. The phenol-furfural compositions are useful for molding large complex forms in which extra time is needed for the resin to fill the mold completely.

**Fabrication and use.** Phenolic resins can be cast from syrupy intermediates that are cured by heating. Laminated products can be produced by impregnating fiber, cloth, wood, and other materials with the resin. After heating, laminated sheets can be pressed into any shapes desired. Most of the phenolic plastics can be machined if necessary.

Cured phenolic plastics are rigid, hard, and resistant to chemicals (except strong alkali) and to heat.

Some of the uses for phenolic resins are for making precisely molded articles, such as telephone parts, strong and durable laminated boards, or for impregnating fabrics, wood, or paper. Phenolic resins are also widely used as adhesives, as the binder for grinding wheels, as ion-exchange resins, and in paints and varnishes. See ADHESIVE; PHENOL; PLASTICS FABRICATION; POLYMERIZATION; TEXTILE CHEMISTRY. [J.A.M.; L.M.H.]

## Phenylalanine



Physical constants of the L isomer at 25°C

$pK_1$  (COOH) 1.83,  $pK_2$  (NH<sub>3</sub><sup>+</sup>) 9.13

Isoelectric point 5.48

Optical rotation  $[\alpha]_D^{25}(\text{H}_2\text{O}) = +34.5$   $[\alpha]_D^{25}(\text{N HClO}) = +4.5$

Solubility (g/100 ml H<sub>2</sub>O) 2.97

Absorption spectrum: peak at 260 mμ (ultraviolet)

An amino acid considered essential for normal growth of animals. The amino acids are characterized physically by the following: (1) the  $pK_1$ , or the dissociation constant of the various titratable groups; (2) the isoelectric point, or pH at which a dipolar ion does not migrate in an electric field; (3) the optical rotation or the rotation imparted to a beam of plane-polarized light (frequently the D line of the sodium spectrum) passing through 1 decimeter of a solution of 100 grams in 100 ml; (4) solubility; (5) absorption spectrum or the wavelength at which maximum absorption occurs. See EQUILIBRIUM, IONIC; ISOELECTRIC

POINT; OPTICAL ACTIVITY; SPECTROPHOTOMETRIC ANALYSIS.

Dietary phenylalanine is the source of tyrosine in animal tissues. Phenylalanine originates, biosynthetically, from phosphoenolpyruvic acid and D-erythrose-4-phosphate, by way of shikimic acid and prephenic acid (see AMINO ACIDS).

During metabolic degradation, the major pathway in mammals is by oxidation to tyrosine, which is then degraded to fumarate and acetoacetate (see TYROSINE). Phenylalanine also can be deaminated to phenylpyruvic acid, of which three metabolic products are known: benzoic acid, phenylacetic acid, and phenyllactic acid. [E.A.AD.]

## Phenylpyruvic oligophrenia

A type of mental deficiency caused by an inherited defect in protein metabolism, specifically the metabolism of phenylalanine (see PROTEIN). It is also known as phenylketonuria. The condition is rare, with an incidence of 1 in 25,000 in the general population, and 3 in 2300 mental deficient. However, research has created the possibility of preventing some types of mental deficiency caused by mutant genes, through medical identification of genetic factors in the parents (see HUMAN GENETICS).

Phenylpyruvic mental deficiency is identified by the presence of phenylpyruvic acid in the urine. The addition of a 5% solution of ferric chloride to the urine causes the formation of a characteristic deep green color in the presence of the acid. Post mortem examination of affected cases discloses absence of the liver enzyme responsible for the metabolism of phenylalanine in normal persons (see ENZYME). The concentration of unmetabolized phenylalanine is associated with diminution of activity in the higher mental centers and in permanent intellectual retardation.

It is possible to alter the amount of the excretion of such abnormal metabolites by altering the amount of phenylalanine in the diet. Phenylalanine-free diets have been developed, and their efficacy is under study. There are some indications that if diet control can be applied in early infancy, damage to the nervous system can be halted and a more normal maturation anticipated.

The mental defective of this type is usually blond, with blue eyes, fair skin, signs of eczema, and a typical musty odor of the urine. Frequent bizarre behavior reactions, including withdrawal, fright reaction, negativism, and posturing, are common in some cases. These cases generally show severe mental retardation although occasionally they may be classified as moderate. They rarely benefit from special education in the public schools or from residential facilities.

From the point of view of genetics, the condition is due to a single mutant gene and follows a typical recessive pattern of inheritance. It is estimated that the gene exists as a recessive in 1 in 173 in the general population. The parents are apparently normal heterozygotes, with no overt clinical evidence of the disease, while the child appears as a

homozygote, with obvious mental deficiency and the marked signs of an abnormal substance in the urine. Recent studies of the heterozygous parents have demonstrated abnormal phenylalanine tolerance curves. This allows the possibility of predicting the existence of the recessive gene and the probable occurrence of a homozygote offspring. Such a technique may provide a concrete basis for medical advice which will protect such potential parents from reproducing and thus reduce the incidence of the disease in the general population. See MENTAL DEFICIENCY. [M.G.K.]

## Phlebitis

An inflammation of the wall of a vein, usually associated with a clot formation, or thrombus, in the vessel. Venous thrombosis is of two main types, phlebothrombosis without inflammation, and thrombophlebitis, where inflammation is involved. The causes are not completely understood; nevertheless, predisposing factors include blood stasis as in prolonged bed rest, tissue destruction, cardiac failure, obesity, varicosities, and infections of a local or systemic nature. As a thrombus is formed, no matter what the cause, portions may break off to become emboli. The emboli may lodge in the lungs with catastrophic results.

In some cases, formation of a thrombus produces an inflammatory reaction of the adjacent vessel wall. If auxiliary, or collateral, circulation is established early, it may relieve the congestion and allow drainage of blood from the affected area. In the absence of venous drainage serious effects may ensue including edema, pain, tenderness, fever, and even gangrene. See CIRCULATION; EMBOLUS; GANGRENE; PAIN, DEEP; THROMBOSIS.

Occasionally, phlebitis may develop without a preceding thrombus, as when trauma or infection produces direct inflammation. In any case, phlebitis often drastically alters both blood vessel walls and the circulation. Deformity or destruction of the venous valves occurs commonly and these in turn act as precipitating factors in the formation of varicose veins, thus establishing a vicious cycle of stasis, thrombosis, and inflammation. Elderly individuals are especially susceptible to phlebitis because of the frequency of previously impaired circulation.

Milk leg is the lay term for both phlebitis and lymph vessel inflammation associated with some pregnancies. [E.G.ST.]

## Phlebotomus fever

A mild, insect-borne virus disease occurring commonly in Mediterranean countries and in Russia, China, and India. It is also known as sandfly fever.

The virus is about 25  $m\mu$  in diameter. Antigenic relationship to other arthropod-borne viruses has not been demonstrated.

After the bite of an infected female *Phlebotomus papatasi*, the patient develops headache, malaise, conjunctivitis, nausea, pain, and stiffness. All patients recover. Clinical diagnosis may be confirmed by antibody rises in blood.

In endemic areas, immunizing infections are acquired by most children. Outbreaks, occasionally mistaken for malaria, occur when susceptible adults, such as military troops, enter an endemic area. Prevention is by control of the vector. See ANIMAL VIRUS. [J.L.M.]

*Bibliography:* T. M. Rivers and F. L. Horsfall, Jr. (eds.), *Viral and Rickettsial Infections of Man*, 3d ed., 1959.

## Phloem

The chief food-conducting tissue in vascular plants. Its conducting cells are known as sieve elements, but phloem may also include companion cells, parenchyma cells, fibers, sclereids, rays, and certain other cells. Phloem is spatially associated with xylem, and the two together form the vascular system. Much less is known of phloem than of xylem, partly because of its lesser direct economic importance and partly because the sieve elements function for a short time (commonly one season) and then undergo marked structural and functional changes including crushing and, in woody plants, sloughing off as a result of periderm formation (see PERIDERM). Further, the development of phloem is more complex than that of the xylem and reliable information in this regard is scanty.

**Sieve elements.** Sieve elements differ from parenchyma cells in the structure of their walls and in the unique character of their protoplasts (see CELL PROTOPLAST; PARENCHYMA). Sieve areas, distinctive structures in sieve element walls, are specialized primary pit fields in which the plasmodesmata (fine strands of cytoplasm) become enlarged as connecting strands and encased in elongate callose collars (Fig. 1). The walls of sieve elements often increase in thickness by the deposition of the so-called nacreous thickening. The protoplast of a young immature sieve element can not be distinguished from that of a typical parenchyma cell of the same age. In the course of development into a functioning sieve element, however, several distinctive changes occur. Slime bodies frequently appear and later fuse into an amorphous mass of slime, the nucleus disintegrates (although the nucleolus may escape and retain its identity), the boundary between the protoplasm and the vacuole becomes indistinct, and the protoplasm eventually becomes virtually unstainable by ordinary histological dyes. Plastids producing a carbohydrate of peculiar character may be present. One or more of these features may be absent from the sieve element; the loss of nucleus is the most common feature.

As a sieve element becomes nonfunctioning, its contents disintegrate. The callose first increases in amount and then disappears along with the connecting strands, leaving empty pores in the wall. The element becomes filled with air and may be crushed later by enlargement of surrounding parenchyma cells.

**Sieve cells and sieve-tube members.** Typical sieve cells are long elements in which all the sieve areas are of equal specialization, though sieve

areas may be more numerous in some walls than in others. In contrast, a sieve-tube member has some sieve areas more specialized than others; that is,

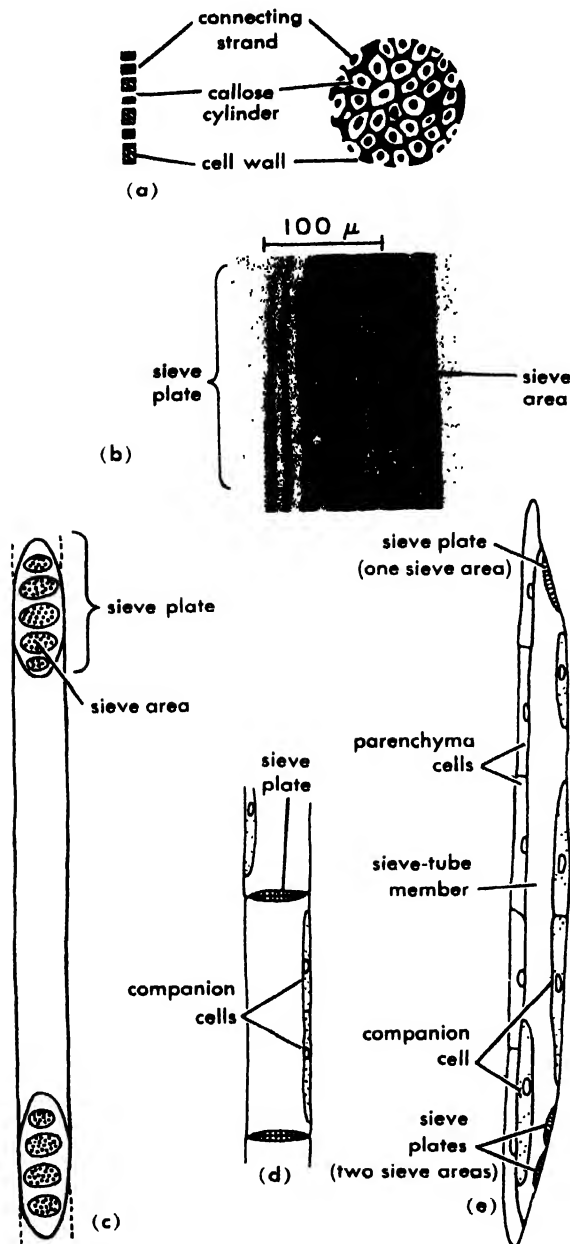


Fig. 1. (a) Part of sieve plate in surface view; connecting strands and wall (black), callose cylinders (clear). (b) Photomicrograph of radial section of part of the secondary phloem of American sycamore (*Platanus occidentalis*) showing the sieve areas on the compound sieve plate of sieve tubes (Forest Products Laboratory, USDA). (c) Oblique sieve plates with several sieve areas in end walls of sieve-tube member. (d) Longitudinal section of a single sieve-tube member and parts of two others; simple sieve plates on transverse end walls and two companion cells. (e) Assemblage of cells derived from one phloem initial: sieve-tube member with sieve plates (hatched) on end walls, strand of parenchyma cells with nuclei, and companion cells (stippled).

the pores, connecting strands, and callose cylinders are larger in some sieve areas. Parts of the walls containing such sieve areas are called sieve plates. Simple sieve plates have one specialized sieve area that generally occurs on a transverse end wall; compound have two or more on an oblique end wall. Sieve tubes are composed of an indeterminate number of sieve-tube members arranged end to end. Thus sieve tubes are to the phloem what vessels are to the xylem. Sieve-tube members are shorter than sieve cells and become progressively more so with evolutionary change. See XYLEM.

**Companion cells.** Companion cells are specialized parenchyma cells that occur in close ontogenetic and physiologic association with sieve-tube members. They arise from the same meristematic cell that produces the sieve-tube member and vary in size, position, and number, but always retain their nucleus. Some sieve-tube members lack companion cells. The precise functional relationship between these two kinds of cells is unknown, but they become nonfunctioning simultaneously.

**Parenchyma cells.** Parenchyma cells in the phloem are thin- or somewhat thick-walled, and occur singly or in strands of two or more cells. They store starch, frequently contain tannins or crystals, commonly enlarge as the sieve elements become obliterated, or may be transformed into sclereids or cork cambium cells. Parenchyma cells in secondary phloem may arise from a meristematic cell (phloem initial) producing only such cells, or from one that also eventually produces one or more sieve-tube members and companion cells. Parenchyma cells seem to intergrade with companion cells in the angiosperms.

**Fibers.** Phloem fibers vary greatly in length (from a fraction of a millimeter in some plants to  $\frac{1}{2}$  meter in the ramie plant). The secondary walls are commonly thick and typically have simple pits.

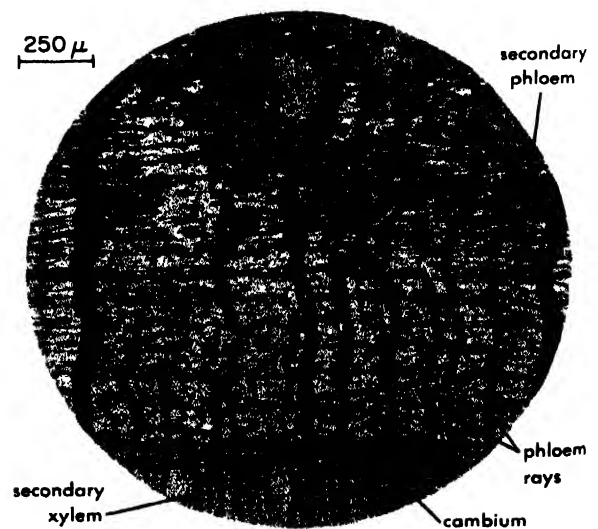


Fig. 2. Photomicrograph of cross section of the secondary phloem of paper birch (*Betula papyrifera*) close to the cambial region. (Forest Products Laboratory, USDA)

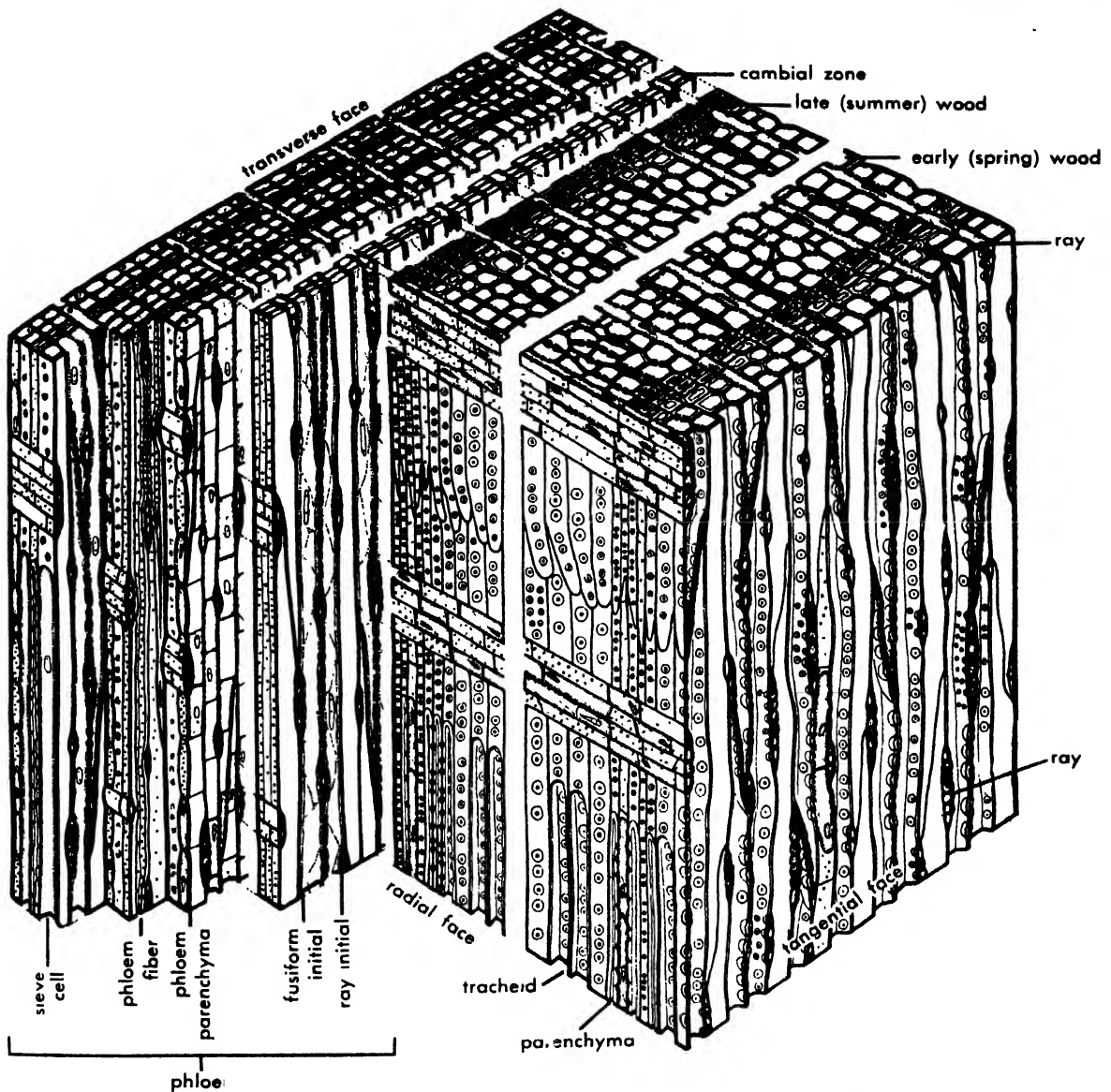


Fig. 3. Block diagram of the secondary xylem, cambial zone, and secondary phloem of the conifer

(gymnosperm) *Thuja* (white cedar). (Courtesy of I. W. Bailey)

but may or may not be lignified. In secondary phloem some fibers do not increase in length beyond the size of their primordia, but others may elongate extensively by apical intrusive growth. In primary phloem, immature fibers elongate, sometimes hundreds of times over their original length. The fibers may become septate, are frequently multinucleate, and may intergrade with sclereids (see **SCLERENCHYMA**).

**Primary phloem.** Primary phloem differentiates from derivatives of the apical meristem (see **MERISTEM, APICAL**). The earliest primary phloem (protophloem) contains chiefly sieve elements, with or without companion cells, and parenchyma cells. The sieve elements function for a brief time and then are obliterated. The remaining cells may become collenchymatous as in many leaves, or be transformed into long protophloem fibers, often erroneously called pericyclic fibers (see **COLLEN-**

**CHYMA; PERICYCLE**). Metaphloem is formed after growth in length of surrounding cells is completed. Sieve elements, companion cells (in angiosperms) and parenchyma cells occur in such phloem, but typical fibers are generally lacking (see **ANGIOSPERMAE**). If secondary phloem is absent, the metaphloem functions throughout the life of the plant.

**Secondary phloem.** Secondary phloem is produced by the same vascular cambium that forms secondary xylem (see **MERISTEM, LATERAL**). Such phloem consists of two interpenetrating systems, the vertical or axial and horizontal or ray (Fig. 2). The phloem rays are basically similar to xylem rays, but their component cells differ in typically lacking secondary walls. Moreover, as the girth of the stem or root increases, the older phloem ray cells increase in width and may divide radially. This dilatation does not occur in all phloem rays, but it is a common feature of secondary phloem

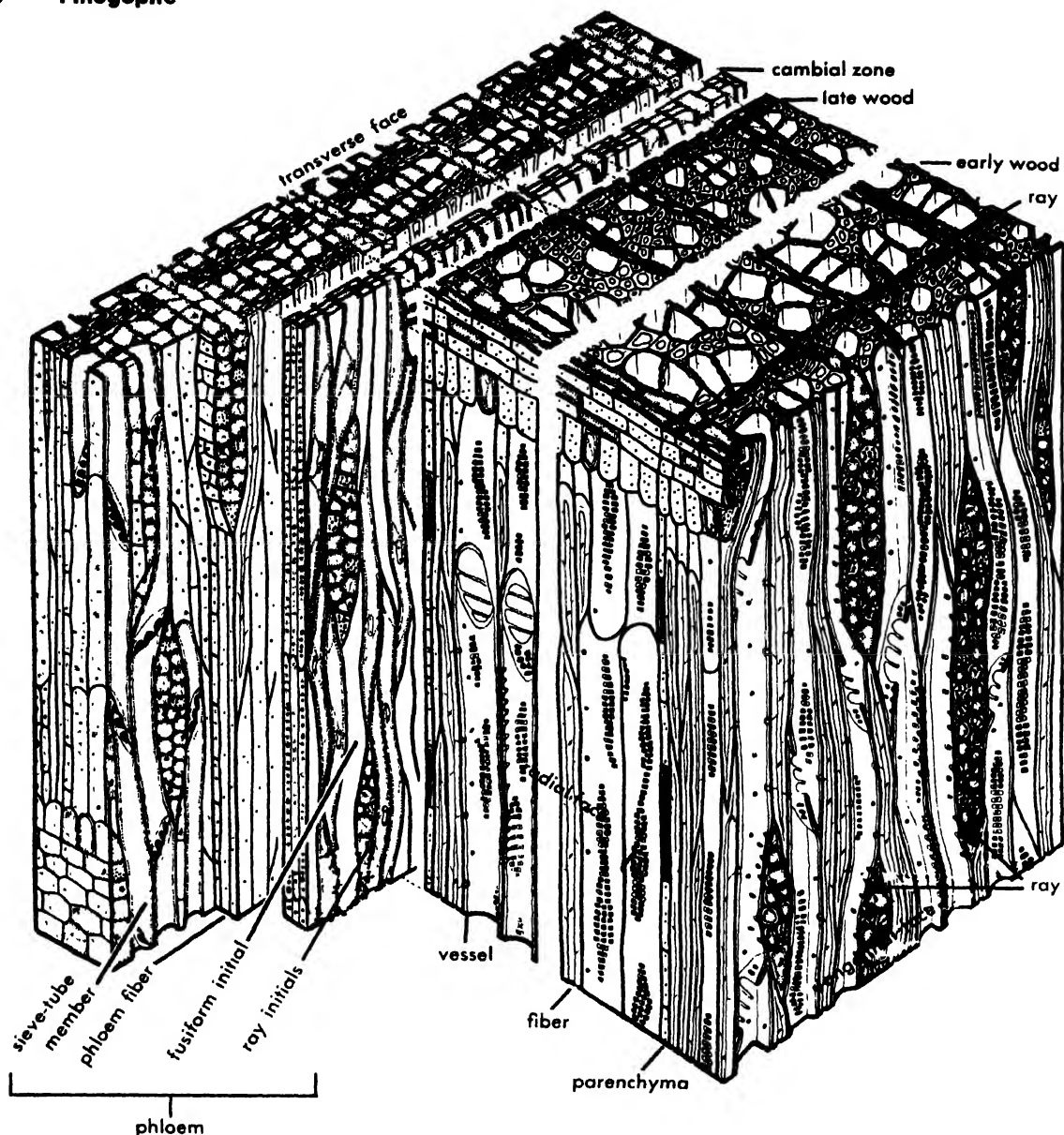


Fig. 4. Block diagram of secondary xylem, cambial zone, and secondary phloem of the dicotyledon

(angiosperm) *Liriodendron* (tulip tree). (Courtesy of I. W. Bailey)

and stops only at the time of periderm formation within the ray. The vertical system contains sieve elements, parenchyma, often fibers or sclereids, and infrequently other elements such as laticifers (see SECRETORY STRUCTURES, PLANT). The fibers may occur singly, in dispersed groups, or in tangential bands.

**Phloem of conifers.** The secondary phloem consists of long sieve cells, parenchyma cells, and frequently of fibers (Fig. 3). These cells may be arranged in regularly alternating bands that give an orderly appearance to the phloem as seen in transsection (see CONIFERALES).

**Phloem of dicotyledons.** The phloem in dicotyledons (Fig. 4) has greater diversity of cell structures and of arrangement than that in the conifers (see DICOTYLEDONEAE). It is composed in varying proportions and groupings of sieve-tube members,

companion cells, parenchyma cells, and often of fibers, sclereids, and various other kinds of cells or cell groups, such as secretory. The various cells may be arranged in alternating bands or have no regular spatial disposition. The functioning phloem is generally more orderly in appearance than the nonfunctioning. This difference results from partial or total collapse of the older sieve elements and associated companion cells, and frequently from the concurrent enlargement of neighboring parenchyma cells. See CYTOLOGY; PLANT TISSUE SYSTEMS.

[V. I. CHEADLE]

*Bibliography:* See PLANT ANATOMY.

## Phlogopite

A mineral of the mica group, also called amber, or bronze, mica. Its composition is  $K_2[Mg, Fe(II)]_6(Si_4, Al_2)O_{20}(OH)_4$ , including minor amounts of



sodium (Na) that substitute for potassium (K) and containing small amounts of Mn, Fe(III), and Ti. With an increase in Fe(II), it grades into biotite from which there is no sharp distinction.

Phlogopite is stable at higher temperatures and has a higher power factor than muscovite, with about the same voltage breakdown. It is widely used as an electrical insulator.

It occurs in disseminated flakes, foliated masses, or large crystals. The basal cleavage is easy and perfect; specific gravity is 2.8-3.0; hardness is 2.5-3.0. Thin sheets are transparent in shades of light brown and green. Reddish-brown reflections are characteristic of cleavage surfaces. It may be colorless to weakly pleochroic.

The structures are monoclinic (one-layer, two-layer, and three-layer pseudorhombohedral). Many phlogopites display asterism in transmitted light because of oriented exsolved rutile needles.

Phlogopite occurs chiefly in certain peridotites (kimberlites), in serpentinized peridotites, in marbles derived from impure dolomitic limestones, and as very large crystals of commercial importance in coarse-grained plagioclase-apatite-calcite-pyroxene rocks of pegmatitic affinity (Ontario and Quebec). See MICA; SILICATE MINERALS. [E.W.H.]

### Phobic reaction

A type of neurosis. The specific forms which phobic reactions, or intense irrational fears whose irrationality the individual may realize without being able to dispel the fear itself, may take are almost as varied as the human imagination. These have often been labeled to render them into a medically acceptable terminology, for example, nyctophobia or morbid fear of darkness; ophophobia, fear of crowds; zoophobia, fear of animals; mysophobia, fear of germs and contamination; claustrophobia, fear of being in a confined space; agorophobia, fear of open places; hydrophobia, fear of water; and syphilophobia, fear of syphilis. These specific fears may be considered neurotic symptoms rather than relatively discrete patterns related to particular psychodynamic foundations, and may occur within the context of broader patterns of maladaptive reaction. The origins of phobic reactions tend to be quite varied; sometimes they originate in a generalized fear of a class of objects, one member of which initially caused a pain reaction. For example, when a child is bitten by a dog it may learn morbidly to fear all dogs. Phobias sometimes result in a more complex, symbolic, and distorted displacement, as when a fear of sexual penetration leads to a fear of all sharp objects.

Counterphobic reactions are the ones in which the person goes out of his way to meet the dangerous object or dangerous impulse head on, as if in an effort to gain mastery by confrontation of the danger. More often than not, this counterphobic maneuver is self-defeating, for the person is directing his efforts, not toward his core problem, but toward some fragmentary representation or some displaced version of it. The daredevil aspects of

such reactions often lead to a kind of heroization of the sufferer, providing a certain amount of secondary gain thereby. See ABNORMAL BEHAVIOR; NEUROSIS. [J.S.B.; W.M.S.]

### Phoebe

Any of three species of flycatchers of the genus *Sayornis*, all found in the United States. The eastern phoebe, *S. phoebe*, is the best known. It nests in southern Canada, throughout the United States east of the Rocky Mountains, and south to northern



The phoebe, *Sayornis phoebe*; length to  $7\frac{1}{4}$  in. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

Georgia. Its wistful call note of "phoe-be," or "pe-wit phoe-be," repeated over and over, and its habit of teetering, or tail-wagging, while sitting on a low perch are readily distinguishing traits. It nests under small bridges and other structures raised a few feet from the ground. See FLYCATCHER; PASSERIFORMES; PEWEE. [J.D.B.]

### Pholidota

An order of mammals which includes only the pangolins or scaly anteaters. These, together with the aardvarks, were long placed in the same order with the South American edentates, until increasing knowledge showed that they are quite distinct. Fossil pangolins are practically unknown, and the ancestry of the group is therefore obscure.

The pangolins are unique among mammals in that they are encased in an armor of overlapping horny scales. Hairs are present on various parts of the body, even between the scales in Asiatic species. Teeth are completely absent, the skull is elongate, and the tongue is worm-shaped, as in the true anteaters. These are adaptations for a diet of termites. Seven species are known from tropical Asia and Africa; all are placed in the genus *Manis*. See EUTHERIA; PHOLIDOTA FOSSILS. [D.D.D.]



## Pholidota fossils

Fossil pangolins are represented by isolated bones from the Pleistocene of Asia, the middle Oligocene of France, and the early Miocene of France and Germany. These few fossils indicate animals similar to living forms: small to medium-sized, large-clawed, quadrupedal, arboreal or terrestrial mammals, with edentulous, tubular skulls well-adapted for a diet of ants. The most remarkable character of the living forms is the complete covering of large, horny, imbricating scales. Pangolin ancestry is largely conjectural because of lack of fossil record, but it may approximate the ancestry of the order Edentata. See EDENTATA FOSSILS.

[D.E.S.]

## Phon

The unit of loudness level. Although the basic unit of loudness is the sone, loudness level is frequently used. The loudness level, in phons, of a sound is numerically equal to the sound pressure level, in decibels, of a 1000-cps reference tone which is judged by listeners to be equally loud, that is, to have the same sone value. The relation between loudness in sones and loudness level in phons is shown in the accompanying nomogram. The usefulness of the phon as a unit is limited by the fact

that the number of phons is not proportional to the subjective loudness of a sound as experienced by a listener.

A major problem in noise control has been the evaluation of the loudness of complex noise. For example, the loudness of a noise which contains energy in a number of octave bands is not equal to the simple sum of the sone values of the individual octave bands present in the noise. The loudness in sones of a complex noise can be estimated with the aid of the following equation:

$$S_T = S_{\max} + 0.3(\Sigma S - S_{\max})$$

where  $S_T$  is the total loudness in sones,  $S_{\max}$  is the number of sones in the loudest octave band, and  $\Sigma S$  is the sum of loudnesses in all the octave bands.

The loudness of a complex sound in sones can be converted to loudness level in phons by use of the nomogram or by the formula

$$\text{Phon} = \frac{1.2 + \log S_T}{0.03}$$

The relation between the subjective loudness of a sound and how noisy it sounds to people has not been determined. See LOUDNESS; see also DECIBEL; SONE. [K.D.K.]

*Bibliography:* S. S. Stevens, Calculation of the loudness of complex noise, *J. Acoust. Soc. Am.*, 28(5):807-832, 1956.

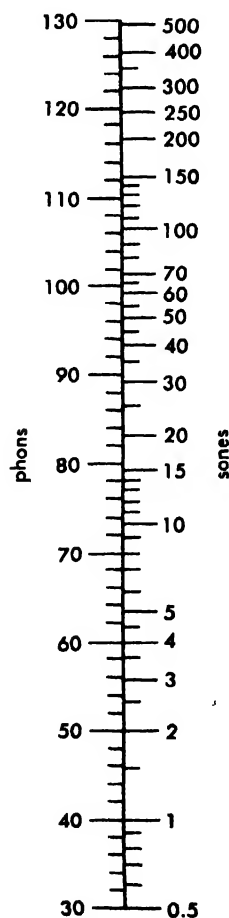
## Phonetics

The study of the sounds which make up speech. It includes the production and reception of the sounds, their classification and listing, their physical characteristics, and the variations of the sounds among languages and dialects.

A set of symbols to stand for the sounds is required in writing. For this purpose, the letters used in English spelling are inadequate. Although they may give some hint as to pronunciation, many letters have a number of different pronunciations, which are indicated in dictionaries by diacritical marks. Other symbols have been adopted by the International Phonetic Association. The table lists 38 of the IPA symbols, with English words for illustration. This list is adequate for most usage in American English.

Some sounds that are thought of as single units are really combinations. Thus the vowel of "my" is a diphthong, in the symbols of the table [maɪ]. The ch of "choice" is [tʃ], and the j of "join" is [dʒ]. On the other hand, some combinations of letters are single sounds, like sh [ʃ], th [θ] [ð], and ng [ŋ].

In a language other than English, some of the sounds in the table would be retained, some dropped, and new ones added. Each symbol of the table represents a phoneme, if it is recognized by those using the language as the same unit, although it may be pronounced in slightly different ways in different combinations or by different people. The slight variations in a phoneme are called allophones. International Phonetic Symbols have been adopted (some as attachments to the symbols in



Nomogram giving the relationship between loudness in sones and loudness level in phons.

## International Phonetic Association symbols

		Vowels	
i	meet	ɔ	all
ɪ	hit	o	note
e	hate	u	boot
ɛ	met	ʊ	foot
æ	hat	ʌ	cup
ɑ	ask	ə	about
ɑ	father	ɜ	bird (Eastern pronunciation)
ɒ	sorry	ɝ	bird (General American pronunciation)
Stop consonants			
p	pit	b	bit
t	to	d	do
k	cap	g	gap
Fricative consonants			
f	fat	v	vat
θ	thin	ð	then
s	see	z	zoo
ʃ	shed	ʒ	vision
h	hat		
Nasal consonants			
m	me	ŋ	sing
n	no		
Glides or semivowels			
w	we	r	red
j	yes	l	let

the table) sufficient to indicate almost all pronunciations in the various languages. A speech utterance written in these symbols is called a phonetic transcription.

The classification used in the table is according to the general method of production of the sounds. The classes may be subdivided, or the sounds may be differently classified such as voiced (having tone from the vocal cords present) or unvoiced (only the frictional noise of exhaled air present). Consonants may be classified according to the location of a constriction, such as bilabial [p,b,m,w], labiodental [f,v], dental [θ,ð], alveolar [t,d,s,z,n,r,l], palatal [ʃ,ʒ,j], velar [k,g,ŋ], and glottal [h].

A speech sound may be analyzed into a frequency spectrum, which shows concentrations of energy in certain frequency regions, called formants of the sound. The technique known as visible speech, showing the smooth transformation of formants from sound to sound, is most enlightening in analysis. See SPEECH. [H.K.D.]

**Bibliography:** C. E. Kantner and R. West, *Phonetics*, 1941; R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech*, 1947.

## Phonocardiography

The science of graphic visualization and interpretation of sound vibrations associated with each heart beat. The phonocardiogram is the record obtained, the phonocardiograph the recording instrument. The latter generally consists of an audio-amplifier coupled to a carbon crystal or capacitance microphone placed over the chest. Band-pass filters may be provided to allow for separation of

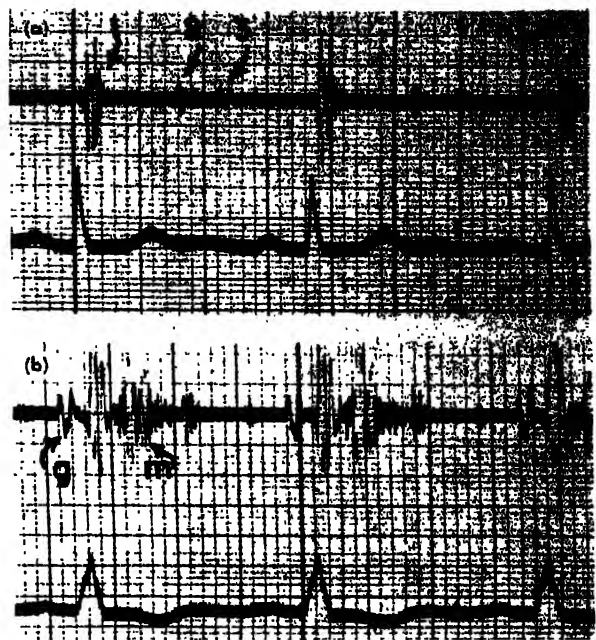


Fig. 1. Phonocardiograms. (a) Normal subject, record from cardiac apex: first (1) and second (2) heart sounds appear split, a faint third (3) heart sound is present. (b) Abnormal phonocardiogram; first and second heart sounds as in (a). The first sound is preceded by an atrial gallop sound (g) coinciding with atrial activity as seen from the electrocardiogram. Between the first and second sound rapid, irregular vibrations occur (m) characteristic of a systolic murmur. Simultaneous electrocardiograms are below each sound record. Time lines equal 0.04 sec.

the auscultatory phenomena of the heart into various frequency ranges.

The phonocardiograms are usually photographic images of galvanometer deflections or oscilloscopic tracings on a time-amplitude scale (see RECORDING INSTRUMENTS, GRAPHIC). For timing purposes phonocardiograms are generally recorded simultaneously with a pulse record, or an electrocardiogram (Fig. 1). Recording techniques generally follow the standard practice of clinical auscultation by comparing records obtained from specified chest positions. For more detailed analysis, phonocardiograms have been obtained from other areas, such as the esophagus, the cavities of the heart, and the large blood vessels by means of special phonocatheters advanced into the heart through the venous system, a procedure known as intracardiac phonocardiography. The miniature microphones placed on the tip of the catheter are barium titanate tubular elements, related to hydrophones used in sonar application (see SONAR). Spectral phonocardiography is an adaptation of the Potter sound spectrograph whereby frequency, time, and loudness of heart noises can be displayed in graphic form (Fig. 2).

Most of the vibrational energy developed during the heartbeat lies below the frequency threshold of the human ear. A small fraction of this energy, however, may be perceived as sound phenomena.

These are of low intensity with fundamental frequencies of 40–200 cycles per second (cps), with occasional higher values. Overtones and harmonics can be heard, but these are attenuated during sound transmission through body tissues and are usually filtered out by the recording devices. For a

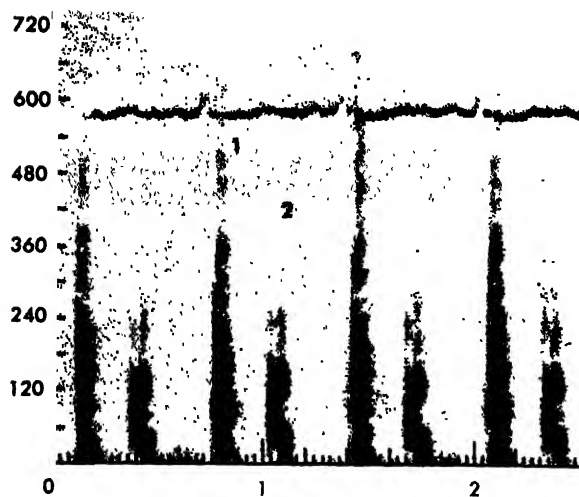


Fig. 2. Spectral phonocardiogram. Frequency-time plot. Intensity of sound proportional to darkness of the tracing. Normal subject, record over region of pulmonary artery. Note splitting of second sound (2). (From V. McKusick, *Cardiovascular Sounds*, Williams and Wilkins, 1958)

sensation of equal loudness the human ear requires greater energy for low than for high frequency sounds. For this reason, almost irrespective of frequency, heart noises remain consistently at the lower threshold of hearing (Fig. 3). In addition, external noises from the room and street decrease sound perception of low intensity sounds, and in heart sounds with mixed frequencies the simultaneous presence of relatively loud low frequency components tends to suppress perception of higher frequency. These phenomena are of obvious clinical importance, and are termed the masking effects. These characteristics of the human ear are not shared by a microphone-amplifier system. Phonocardiograms therefore cannot be considered faithful graphic reproductions of sounds as perceived by a physician through a stethoscope. A logarithmic phonocardiograph has been described which more nearly records the heart sound vibrations as perceived by human subjects.

Clinical practice distinguishes heart sounds from heart murmurs. Physically they are differentiated only by duration because both are irregular sound mixtures of low frequencies and variable intensities. The normal heartbeat is associated with two loud heart sounds occurring at the beginning and at the end of muscular contraction or systole, coincident with the closure of the heart valves. In young subjects a third sound of low intensity is commonly heard shortly after the second sound, and a fourth sound is occasionally present preceding the first

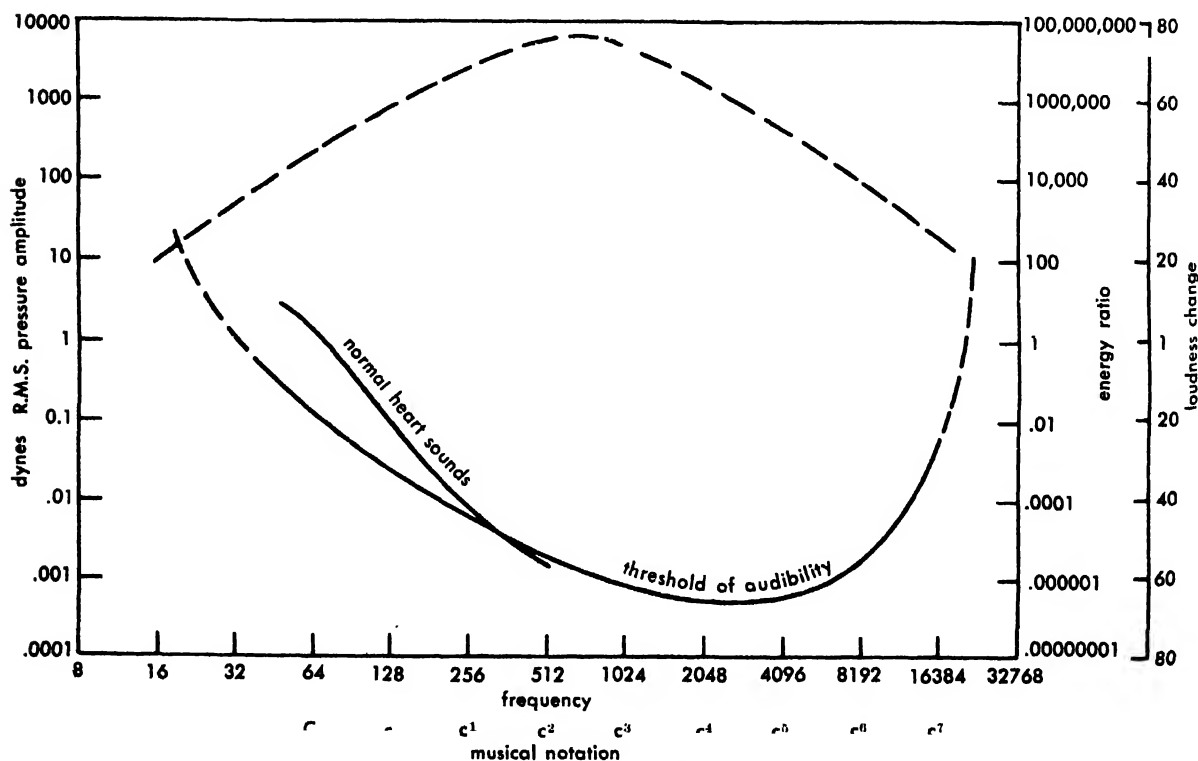


Fig. 3. Normal heart sounds and the characteristics of hearing. The range of human hearing is indicated by the curved lines with higher sound intensities required at lower than at higher frequencies to give the

sensation of equal loudness. Intensity and pitch of heart sounds always remain close to the threshold of audibility. (From H. B. Williams and H. F. Dodge, *Arch. Internal Med.*, 38:685, 1926)

sound and related to the contraction of the atria. These additional sounds may become accentuated in heart failure at any age and, together with the first and second sounds, may give the appearance of triple or quadruple rhythm known as gallop rhythms. Asynchronous ventricular contraction, valve narrowing, and extracardiac adhesion may give rise to duplication of heart sounds, clicks, and scratches of brief duration. Heart murmurs may occur as the consequence of rapid bloodflow in an otherwise normal heart. These are known as flow murmurs or ejection murmurs, or they may be associated with valvular and other structural intracardiac defects. The technique of clinical auscultation supplemented by phonocardiography is concerned with identifying pitch (frequency), intensity, timing, duration, and maximum area of transmission of sounds and murmurs with specific structural changes within and about the heart.

The physical basis for the origin of heart sounds and murmurs is largely speculative. In a closed fluid compartment, such as the vascular system, the release of sound energy requires turbulent flow with eddy or vortex formation caused either by excessive speed of flow or by normal or pathologic obstructions resulting in sudden changes in blood velocity. The occurrence of high frequency transients, and of murmurs with tonal (musical) quality suggests that under pathologic circumstances solid structures within the heart may be set in vibration giving rise to abnormal noises. The formation and collapse of vapor or gas pockets (cavitation) occurring in local areas of high velocity has also been claimed as a source for pathologic sounds. See BIOPHYSICS; ELECTROPHYSIOLOGY (HEART). [H.HE.]

**Bibliography:** V. McKusick, *Cardiovascular Sounds*, 1958; M. B. Rappaport, H. B. Sprague, The graphic registration of the normal heart sounds, *Am. Heart J.*, 23:591, 1942; S. Rodbard (ed.), *Symposium: Present Status of Heart Sound Production and Recording*, IRE, Trans. on Med. Electronics, PGME-9, December, 1957; H. B. Williams and H. F. Dodge, Analysis of heart sounds, *Arch. Internal Med.*, 38:685, 1926.

## Phonograph

An instrument for recording (reproducing) acoustical signals, such as voice and music, by transmission of vibrations from (to) a stylus that is in contact with a groove in a rotating disk. For an extended discussion of phonographs, phonograph records, and related topics, see DISK RECORDING.

[H.F.O.]

## Phonolite

A light-colored, aphanitic (not visibly crystalline) rock of volcanic origin, composed largely of alkali feldspar, feldspathoids (nepheline, leucite, sodalite), and smaller amounts of dark-colored (mafic) minerals (biotite, soda amphibole, and soda pyroxene). Phonolite is chemically the effusive equivalent of nepheline syenite and similar rocks. Rocks

in which plagioclase (oligoclase or andesine) exceeds alkali feldspar are rare and may be called feldspathoidal latite. See FELDSPATHOID.

Rapid cooling at the surface causes lavas to solidify with very fine-grained textures. Most phonolitic lavas, however, carry abundant large crystals (phenocrysts) when they are erupted, and these are soon frozen into the dense matrix to give a porphyritic texture. Generally very little material congeals as glass. The phenocrysts, many visible to the naked eye, include alkali feldspar, feldspathoids, and mafics. These may be well-formed (euhedral) or moderately well-formed (subhedral).

Most other features of phonolites can be seen only microscopically. The alkali feldspar is principally soda-rich sanidine and orthoclase. It generally occurs in the rock matrix, but if abundant it may also form as phenocrysts. Plagioclase is not abundant except in nepheline latites where it may form abundant phenocrysts.

Nepheline may occur as euhedral crystals (square or hexagonal) some of which may be phenocrysts. Otherwise it is irregular (anhedral) and interstitial. Nosean, hauyne, and sodalite, as euhedral or partly corroded crystals, may occur as phenocrysts and matrix grains. These twelve-sided (dodecahedral) crystals generally show hexagonal outlines in thin sections of the rock. Eight-sided, euhedral crystals of pseudoleucite may occur as phenocrysts in potash-rich rocks. More rounded grains of leucite may form part of the matrix. Leucite is commonly altered to pseudoleucite, but the euhedral outline is retained. Analcite occurs principally as matrix material but in some rocks it is abundant and as large euhedral phenocrysts.

Biotite is not common but may form large strongly resorbed phenocrysts. Amphiboles are usually soda-rich (riebeckite, hastingsite, and arfvedsonite). They may occur as phenocrysts or as interstitial clusters. They may show resorption or may be replaced by pyroxene. The most important mafic is soda pyroxene. As phenocrysts it is commonly zoned with cores of diopside surrounded by progressively more sodic shells of aegirine-augite and aegirite. Aegirite is the common pyroxene of the rock matrix.

Accessory minerals are varied and include sphene, magnetite, zircon, and apatite.

The structures and textures of phonolite are similar to those of the more common rock trachyte. Fluidal structure, formed by flowage of solidifying lava and expressed by lines or trails of phenocrysts, may be seen without magnification. Under the microscope, flowage is shown by subparallel arrangement of elongate feldspar crystals. See TRACHYTE.

Phonolites are rare and highly variable rocks. They occur as volcanic flows and tuffs and as small intrusive bodies (dikes and sills). They are associated with trachytes and a wide variety of feldspathoidal rocks.

The origin of phonolites and related rocks constitutes an interesting problem. There is still considerable difference of opinion as to how the phono-

litic magma (molten material) originates. One theory assumes an origin from basaltic magma by differentiation. Certain early formed crystals are removed (perhaps by settling) causing the residual magma to approach the composition of phonolite. Another theory supposes these peculiar magmas to form when a more normal rock melt assimilates large quantities of limestone fragments. Volatiles, notably carbon dioxide, are considered by many to play an important role in transferring and concentrating certain constituents (like potassium) in the magma. The great variety of rock types and modes of association strongly suggests that several different mechanisms may operate to form these feldspathoidal rocks. See IGNEOUS ROCKS; MAGMA.

[C.A.C.A.]

## Phonon

A sound quantum. The energy of a phonon is  $h\nu$ , where  $h$  is Planck's constant and  $\nu$  the frequency of vibration of the sound wave. The phonon is thus analogous to the photon, a light quantum.

In treatments of the scattering of electrons and other particles by thermal waves (short sound waves) in matter, the selection rules which arise bear a formal resemblance to the laws of conservation of energy and momentum holding for collisions between particles. This leads to the concept of a phonon as a packet of sound waves, the wave packet having particle-like aspects. The concept is particularly convenient in the theory of the thermal conductivity of insulators, where one may speak of a phonon gas, collisions between phonons, and a phonon mean free path. In the theory of the properties of superfluid helium, the quanta of longitudinal sound waves in the liquid helium are called phonons. See CONDUCTION (HEAT).

[J.D.L.]

## Phonoreception

The perception of sound by animals through specialized sense organs. A sense of hearing is possessed by animals belonging to two divisions of the animal kingdom, the vertebrates, which form the main subphylum of the phylum Chordata, and the insects, which comprise the most important class of the phylum Arthropoda. This sense is mediated by the ear, a specialized organ for the reception of vibratory stimuli. Such an organ is found in all except the most primitive vertebrates, but only in some of the many species of insects. The vertebrate and insect types of ear differ in evolutionary origin and in their modes of operation, but both have attained high levels of performance in the reception and discrimination of sounds. See SOUND.

**Vertebrates.** The vertebrate ear is a part of the labyrinth, located deep in the bone or cartilage of the head, one on either side of the brain. A complex assembly of tubes and chambers contains a membranous structure which bears within it a number of sensory endings of different kinds.

The membranous labyrinth is shown in a generalized schematic form in Fig. 1. It is convenient to recognize two divisions, a superior division, which

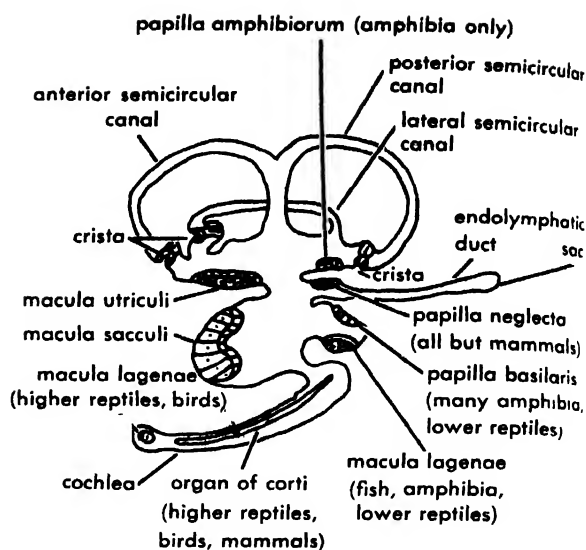


Fig. 1. Generalized sketch of the vertebrate labyrinth. The three cristae, macula utriculi, and macula sacculi are always present in vertebrates, and the other endings appear as indicated, with a few exceptions.

includes the three semicircular canals and the utricle, and an inferior division, which includes the saccule and its appendages, the lagena and the cochlea. The superior division is remarkably uniform in character from the higher fishes upward, but the inferior division shows many variations. The saccule is always present. The lagena is present in all classes except the mammals, although it is missing in occasional species. The cochlea is found in reptiles, birds, and mammals.

The sensory endings within these parts of the labyrinth also vary in the vertebrate series. Again there is uniformity for the superior division. There is a crista in each ampulla of the three semicircular canals and a utricular macula. In all but the mammals (with a few individual exceptions), there is a macula neglecta, usually located on the floor of the utricle or close to the junction of utricle and saccule. All forms have a saccular macula. All those with a lagena (in general, all except the mammals) have a lagenar macula. All the amphibians have a papilla amphibiorum, but it is found in no other forms. A basilar papilla appears in certain amphibians, is continued in the reptiles, and then is developed in a more elaborate form as the cochlea of higher reptiles, birds, and mammals.

These endings contain ciliated cells (hair cells) which are supplied by fibers of the eighth cranial (auditory) nerve. In the cristae the cilia of the hair cells are particularly long and are embedded in a gelatinous substance that forms a cap or cupola. In the maculae the cilia are surmounted by a flat plate of gelatinous material in which numerous granules of calcium carbonate (otoliths) are usually embedded. The ciliated cells in the papillae lie on a movable membrane (basilar membrane) and have a membranous covering, the tectorial membrane.

The superior part of the labyrinth generally serves for bodily posture and equilibrium, whereas the saccule and its appendages (lagena, cochlea) serve for hearing. However, there are exceptions to this rule, the most important of which is that in the higher vertebrates, including mammals and probably birds and reptiles, the saccule serves only for equilibrium. See EQUILIBRIUM, BIOLOGICAL.

Beginning with the amphibians, which are the earliest vertebrates to spend a considerable portion of their lives on land, there appears a special mechanism, the middle ear, whose function is the transmission of aerial vibrations to the endings of the inner ear. All the vertebrates above the fishes, and certain of the fishes as well, have some type of sound-facilitative mechanism.

**Fishes.** The maculae of the utricle, saccule, and lagena in the bony fishes have a peculiar form. Instead of numerous calcareous particles there is a single otolith, a large body of distinctive form. The macula neglecta is sometimes lacking.

Few questions have been more actively debated than the ability of fish to hear. Experiments on this question began with G. Parker in 1903, who observed the natural reactions of fish when exposed to a sudden sound, and were carried forward by F. Westerfield and others in 1922 by the introduction of conditioned-response methods. This work culminated in the series of studies by K. von Frisch and his associates, who trained fish to make feeding responses at the sounding of a tone. These experiments proved that fish may be divided into two groups according to hearing ability, those that hear only crudely and those that hear well. The first group includes the great majority of fish species, with what may be called the basic type of labyrinth and lacking any accessory mechanism. The fish that hear well have one of two general types of sound-facilitating structure: either an air vesicle adjacent to some part of the labyrinth or a connection with the swimbladder.

The second of these types of accessory structure is the more common, and is found in a large group of fresh-water fishes known as the Ostariophysi. Between the labyrinth and the anterior part of the swimbladder is a chain of three or four small bones, known from their discoverer, E. H. Weber, as the Weberian ossicles (Fig. 2). Weber correctly supposed, when he described this apparatus in 1820, that it serves for the facilitation of sound reception, for it has been demonstrated that the hearing is impaired after removal of the swimbladder and after an interruption of the ossicular chain.

The most extensive study of hearing in fish has been made on the European minnow (*Phoxinus laevis*), one of the Ostariophysi that responds readily to training procedures. This fish is able to hear tones over a range from 32 cycles per second (cps) or a little below to 5000 or 6000 cps, and in the lower part of this range can discriminate a change of frequency of about 3%. The dwarf catfish, *Ictalurus (Ameiurus) nebulosus*, responds to tones as high as 13,000 cps.

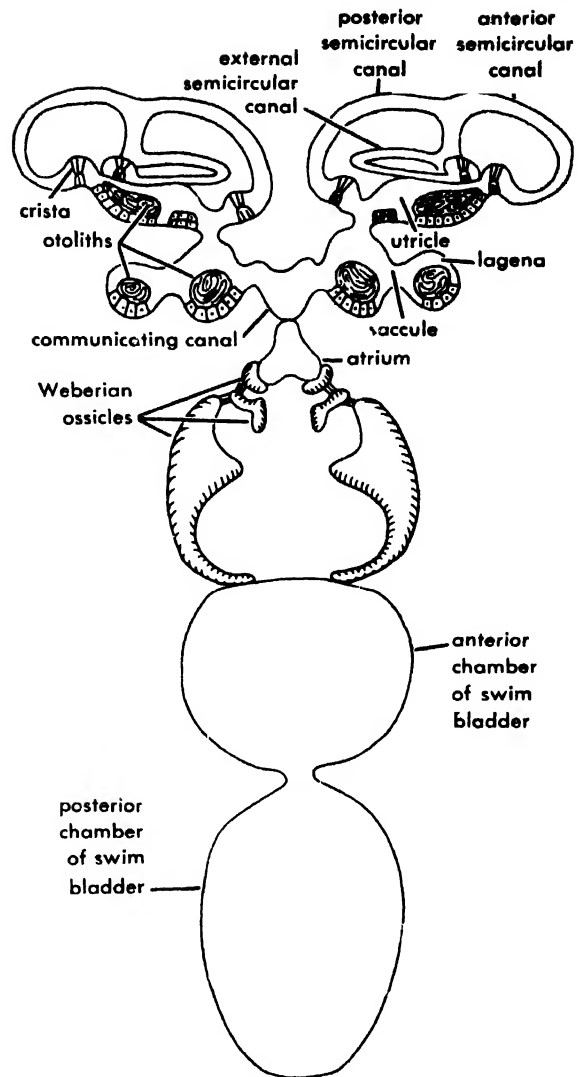


Fig. 2. The two labyrinths of an ostariophysan fish and their connections with the swimbladder through the Weberian ossicles. Diagrammatic view, from above.

For morayrids and labyrinthine fishes, which are forms with an air vesicle adjacent to the labyrinth, the upper limit of hearing is around 3000 cps. In ordinary fish (which lack the accessory mechanism) this limit is usually found between 600 and 800 cps, and the sensitivity is 30–40 decibels (db) poorer than in the Ostariophysi.

A number of experiments, mostly on minnows, have dealt with the problem of the particular parts of the fish labyrinth that are concerned with hearing. Removal of the superior portion, which includes the utricle and semicircular canals, does not impair the responses to sound, but seriously affects the posture and swimming ability. After this operation the fish may assume an inverted position, and swims erratically. These parts must therefore contain organs of equilibrium. Removal or impairment of either the saccule alone or the lagena alone leaves the fish able to hear, but the removal of both saccule and lagena abolishes the responses to all tones except those of very low frequency which are

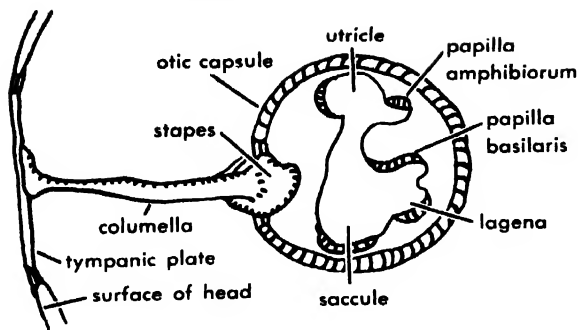


Fig. 3. Ear of the frog. Diagrammatic, simplified.

perceived through skin receptors. Hence, the endings of the sacculus and lagena are auditory in function. Exceptional in this respect are the herring and sardine (clupeids), in which an air vesicle is applied to the wall of the utricle; in these the sense of hearing is probably mediated by the utricular macula.

**Amphibians.** The three orders of amphibians—the Apoda (legless), including wormlike forms such as caecilians; the Caudata (tailed), including mud puppies, newts, and salamanders; and the Salientia (tailless), including frogs and toads—all have some type of middle-ear mechanism.

The first two orders include animals whose ears show a great variety of accessory structures, some of which look as though they might function well in sound reception, whereas others seem crude. Of these only the salamander has been studied experimentally. In 1938, S. Ferhat-Akat trained larvae to come for food at the sounding of a tone, and got results for tones up to 244 cps in one specimen and up to 218 cps in three others.

Higher amphibians, such as the frog, possess a well-developed middle-ear mechanism, consisting of a disk of cartilage flush with the lateral surface of the head and covered with skin, and a rod of cartilage and bone, called the columella, leading inward from the disk and expanding to form the stapes which is imbedded in an opening (oval window) in the wall of the otic capsule (Fig. 3).

The active and often loud croaking of frogs in the breeding season has focused attention upon the problem of their hearing. R. Yerkes in 1905 first succeeded in obtaining experimental evidence of their auditory sensitivity by showing that sounds may enhance or inhibit their response to a strong tactual stimulus. Several studies in which the impulses from the eighth nerve were recorded on stimulation with tones showed results only for low frequencies, up to 500 or 600 cps, or at most to 1024 cps. The most extensive study of the electrical responses of the ear by W. Strother in 1958 showed responses in *Rana catesbeiana* over a range from below 100 cps to about 3500 cps. The sensitivity was best at 400–1500 cps, and then fell off rapidly to the upper limit.

**Reptiles.** The living reptiles belong to four important groups, represented by snakes, turtles, chameleons and lizards, and crocodiles and alligators.

Many authorities have asserted that snakes are completely deaf, or that their ears are sensitive only to vibrations conducted to the head through the ground. This impression has arisen partly from the fact that snakes do not have any external ear and do not show obvious reactions to sounds. There is no tympanic membrane to receive aerial sound pressures, but its purpose is served by one of the bones of the skull, the quadrate bone, which is loosely attached to the main part of the skull. Although it lies beneath the skin and other tissues of the side of the head, the quadrate bone presents a flat surface for the action of sounds, and communicates them to a thin bony rod (columella) running inward to expand as the stapes in the oval window (Fig. 4).

Recent experiments have shown that electrical potentials are produced in the inner ears of snakes in response to low-frequency sounds, to both the sounds conducted through the substratum and those conducted through the air in the usual way. Hence, it seems safe to conclude that snakes have hearing, although only for the lower range of sounds and not as highly sensitive as that of most other animals.

Doubt has often also been expressed about the ability of turtles to hear, but here again the evidence is that they do. They have a well-developed middle ear, including a cartilaginous disk on the side of the head beneath the skin, and a columella leading to a stapes in the oval window of the otic capsule. Two investigators have succeeded in training turtles to make positive reactions to an acoustic signal, although others have failed in this attempt. The electrophysiological method yields positive results. Indeed, the observations show that for low tones, those of 100–700 cps, the turtles have excellent sensitivity, with the wood turtle, *Clemmys insculpta*, exceeding other species studied (Fig. 5).

Structurally the ear of the lizard is superior to that of the turtle. There is a membranous drum, a columella, and a stapes. A few studies have dealt with their hearing, and there is no doubt that they hear sounds in the middle range of frequencies.

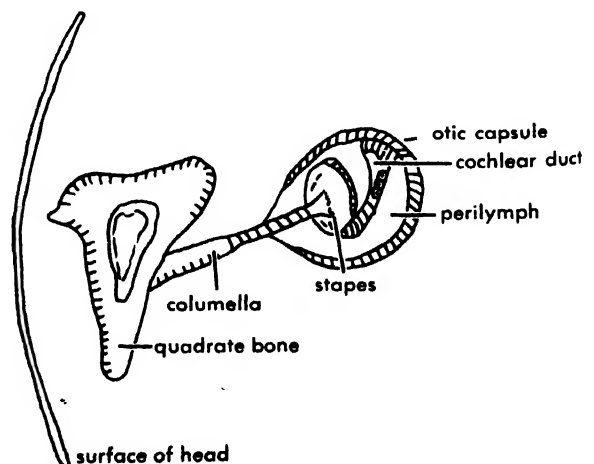


Fig. 4. Diagram of the ear of a snake. The nonauditory parts and endings are not shown.



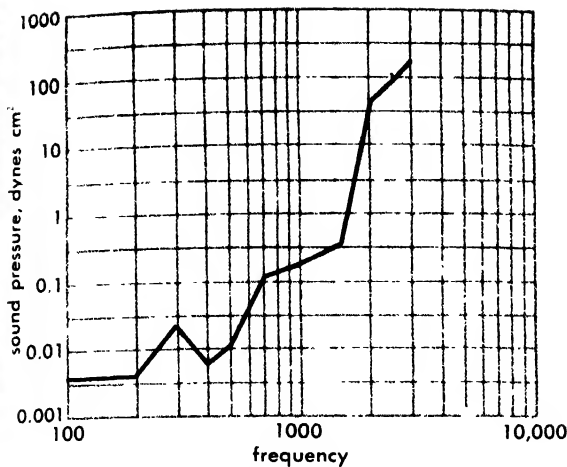


Fig. 5. The auditory sensitivity of a wood turtle, *Clemmys insculpta*, as shown by the potentials produced in its ear by sounds. The curve shows the sound pressure necessary to produce a potential of 0.3 microvolts.

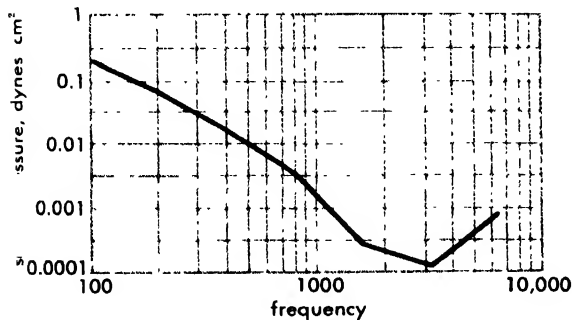


Fig. 6. Threshold sensitivity in the bullfinch (mean of 4 birds). (From J. Schwartzkopff, *Über Sitz und Leistung von Gehör und Vibrationssinn bei Vögeln*, Z. vergl. Physiol., 31:527-608, 1949)

C. Berger was able to train two species of lizards (*Lacerta agilis* and *L. vivipara*) to make feeding movements in response to various sounds, including tones over a range from 69 to 8200 cps. Electrical potentials have been recorded from the ears of the lizard *Anolis* for tones in the range of 100-10,000 cps.

The crocodiles have an ear representing a distinct advance over other reptiles, and there is a curved cochlea similar to that of birds. Probably they have excellent hearing, although the experimental evidence is scanty as yet. Many general observations indicate that they use sounds in mating activities, and the males are capable of producing loud roars. F. Beach was able to provoke captive animals into roaring and making movements by stimulation with low tones, especially sounds at 57 cps, but also others at about 300 cps. E. Wever and J. Vernon found that young caimans gave cochlear potentials in response to sounds over a range of 20-6000 cps.

**Birds.** The labyrinth in birds is generally similar to that of the higher reptiles. There is a membranous eardrum, a columella leading inward to the

stapes, and a curved cochlea. There are only minor variations in the forms of these structures among the various species.

Most birds have a range of hearing of about 50-20,000 cps, and an absolute sensitivity that probably approaches that of man in the medium-high tone range. The threshold sensitivity of a finch, *Pyrrhula p. minor*, is shown in Fig. 6, as determined by J. Schwartzkopff by a training method. In general, the small songbirds are more sensitive than the larger birds such as chickens and pigeons. The owl is exceptional in having an ear with a particularly large drum membrane and other special features that make for high sensitivity. Thus, an owl perched in a tree at dusk is able to hear and to locate a mouse rustling in the grass below.

Pitch discrimination in songbirds and parrots is about as keen as that of man (0.3-0.7%) but in pigeons it is relatively poor (6%).

**Mammals.** The auditory apparatus attains its highest development in the mammals. The columella of lower forms has been replaced by a chain of three ossicles, which connect the tympanic membrane with the inner ear. In the egg-laying mammals, such as the platypus, the cochlea is a curved tube as in crocodiles and birds, but in all other mammals it is a spiral of 1-4 turns. The great extension of the cochlea and the corresponding multiplication of sensory cells have enhanced the capacities of the mammals to deal with the varieties and complexities of sounds.

Despite intense interest in mammalian hearing, precise information is available on only a few mammals apart from man himself. Experimental studies have been carried out on some of the subhuman primates and on a few of the common laboratory animals. Only fragmentary information is available on the many other species of mammals, although it can be assumed from their general behavior, because they seem to respond to much the same range and intensities of sounds that man does, that their hearing is similar to man's.

Training experiments are easily carried out on the subhuman primates, and Fig. 7 presents threshold curves for the chimpanzee, the rhesus monkey, and a species of marmoset, with the human curve for comparison. It will be noted that these animals have auditory sensitivity similar to man's in the low-tone range, but are superior to man in the high-tone range.

The most extensive studies have been made on the cat. Its sensitivity also is similar to man's over the lower range, but extends far above the human limit, to 60,000 cps or more. Electrical potentials have been recorded from the cat's cochlea to tones as high as 100,000 cps.

Other animals whose hearing has been studied experimentally are the dog, rat, guinea pig, rabbit, certain species of mice, and bats.

The bats are of special interest, more particularly the small insectivorous species, because they repeatedly produce vocal sounds of high frequency, up to 40,000 cps, as they fly about in search of insect prey. They locate the prey by hearing the

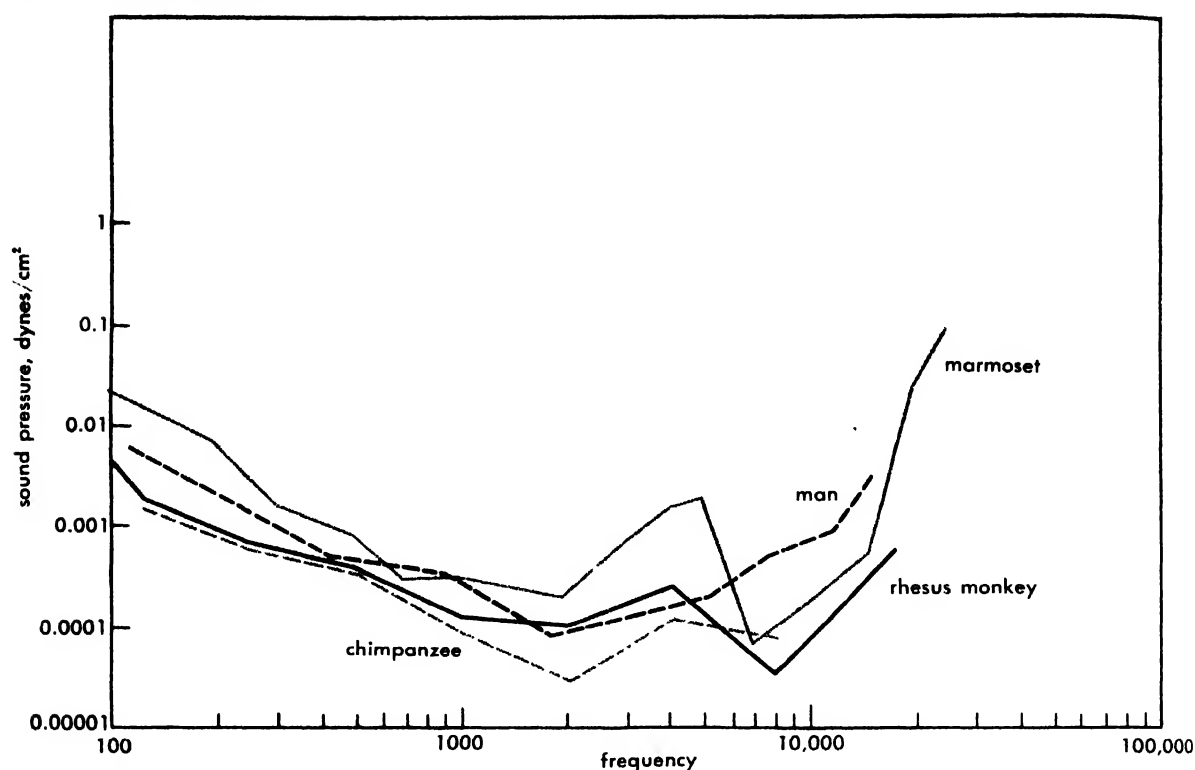


Fig. 7. Auditory thresholds in primates. The curves show the sound pressures that are barely audible for man, chimpanzees, rhesus monkeys, and marmosets.

(Data from Sivian and White, Elder, Harris, Wendt, and Seiden)

echoes of their cries. Similarly they guide themselves in the total darkness of caves by echoes from the walls. The hearing of bats extends far into the high frequencies, up to 100,000 cps at least. See EAR; HEARING.

**Invertebrates.** The group of invertebrates which has received the most attention has been the insects. Other arthropods such as certain crustaceans and spiders have also been found to be sensitive to sound waves.

**Insects.** The insect ear consists of a superficial membrane of thin chitin with an associated group of sensillae called scolophores. Such an apparatus is shown in simplified form in Fig. 8. These ears are found in most species of katydids, crickets, grasshoppers, cicadas, waterboatmen, mosquitoes, and nocturnal and spinner moths. They occur in different places in the body: on the antennae of mosquitoes, on the forelegs of katydids and crickets, on the metathorax of cicadas and waterboatmen, and on the abdomen of grasshoppers. Probably these differently situated organs represent separate evolutionary developments, through the association of a thinned-out region of the body wall with sensillae that are found extensively in the bodies of insects and by themselves seem to serve for movement perception.

The insects mentioned above are noted for their production of stridulatory sounds made by rubbing the edges of the wings together, or a leg against a wing. These sounds are produced by the males and serve for enticing the females in mating.

The sensitivity of insect ears is keenest in the high frequencies. Figure 9 shows threshold curves obtained on a katydid by observing the potential-produced in the auditory nerve during stimulation with sounds. As will be seen, the sensitivity in this species is greatest in the region of 7000–60,000 cps. It extends to even higher frequencies, usually as high as 120,000 cps and sometimes beyond. Other species have distinctly different sensitivity curves, and there is reason to believe that there is a relation to the range of the stridulation sounds. There is evidence that these sounds are discriminated, for

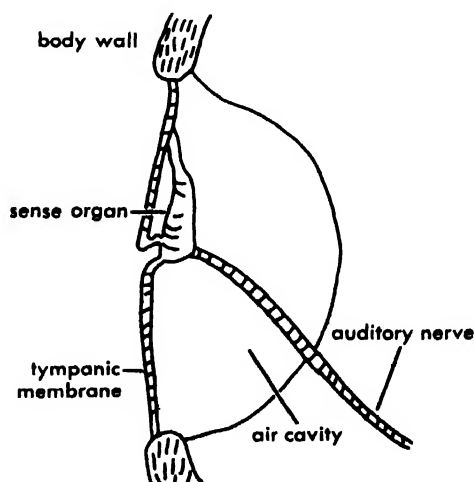


Fig. 8. Ear of a grasshopper; diagrammatic.

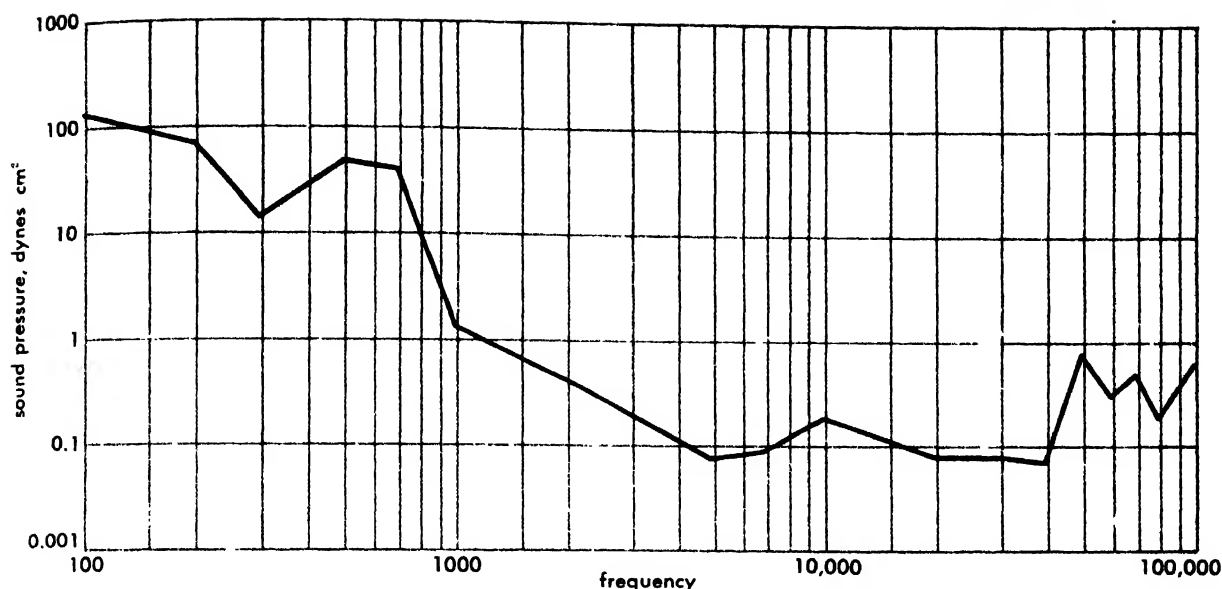


Fig. 9. Auditory thresholds of the katydid, *Conocephalus strictus*. (From E. G. Wever and J. A. Vernon,

*The auditory sensitivity of Orthoptera*, *Proc. Natl. Acad. Sci.*, 45:413-419, 1959)

the females of one species respond to the males of their own kind. J. Regen placed a cage of chirping male crickets in a field, and found that females of their species sought them out. This seeking activity ceased, however, when the males were deprived of their stridulating organs or the females were made deaf by removal of their ears. A most striking adaptation is that shown by mosquitoes: the ear of the male mosquito is sensitive only to a narrow range of frequencies around 380 cps, and this frequency is the one produced by the wings of the female in flight. M. Tischner found that when the ear of a male mosquito was made nonfunctional, the mosquito failed to find a mate.

[E.G.W.]

**Bibliography:** F. Eggers, *Die stiftführenden Sinnesorgane*, 1928; K. von Frisch, *Über den Gehörsinn der Fische*, *Biol. Revs.*, 11(27):210-246, 1936; D. R. Griffin, *Listening in the Dark*, 1958; J. Schwartzkopf, *Über Sitz und Leistung von Gehör und Vibrationssinn bei Vögeln*, *Z. vergl. Physiol.*, 31:527-608, 1949; E. G. Wever and J. A. Vernon, *The auditory sensitivity of Orthoptera*, *Proc. Natl. Acad. Sci.*, 45:413-419, 1959.

## Phoresy

A relationship between two different species of organisms in which the larger, or host, organism transports a smaller organism, the guest. It is regarded as a type of commensalism in which the relationship is limited to transportation of the guest. The term is credited to P. Lesne following his observations on the biology of a small fly, *Limosina sacra*, which is transported by a scarabeid, one of the dung beetles, into its burrow. These burrows are suitable breeding sites for both animals.

[C.B.C.]

**Bibliography:** P. Lesne, *Moers du Limosina sacra*, Meig. *Phénomènes de transport mutuel chez*

les Animaux articles. *Origines du parasitisme chez les Insectes dipteres*, *Bull. Soc. Ent. France*, 162-165, 1896.

## Phoronida

A small, relatively homogeneous group of animals now generally considered to constitute a separate animal phylum, although in the past they have been grouped with other phyla such as the Annelida, Molluscoidea, and Chordata. Two genera, *Phoronis* and *Phoronopsis*, and about sixteen species are recognized at the present time; however, the taxonomy of the group is in need of thorough revision.

**Habitat and distribution.** Phoronids may occur in vertical tubes placed just below the surface in intertidal or subtidal mud flats or as feltlike masses of intertwined tubes attached to rocks, pilings, or old logs in shallow water. In both cases the tubes, composed basically of a secreted, parchmentlike material, are encrusted with small particles of sand or shell. A third living habit concerns those phoronids found inside channels, probably self-made, in limestone rock or the shells of dead pelecypod mollusks.

The geographical distribution of phoronids appears to be world-wide in temperate and tropical seas. There are no records from the polar regions.

**Morphology.** The body is more or less elongate, ranging in length from about 4-20 cm, and bears a crown of tentacles arranged in a double row surrounding the mouth which is usually crescent-shaped. The anus occurs at the level of the mouth and is borne on a papilla immediately outside the double row of tentacles. The digestive tract is therefore U-shaped, the mouth and anus opening close together at one end of the animal. The tentacles rest on a connective tissue base known as the lophophore (see **LOPHOPHORE**). The double row of

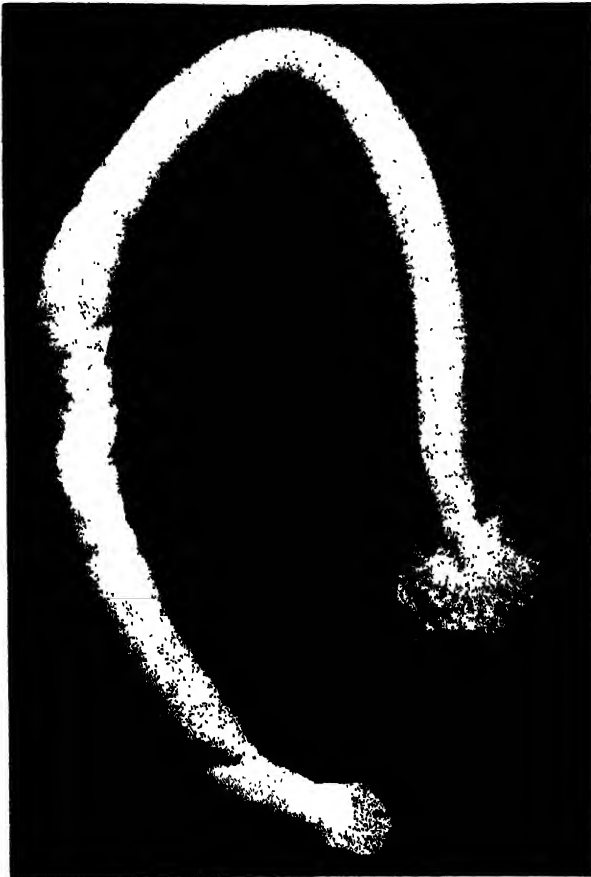


Fig. 1. *Phoronopsis harmeri* removed from its tube. Length about 20 cm.

tentacles may form either a slightly indented circle or a complex double spiral. The tentacles vary in number from about 50 to over 300, are ciliated, and create a feeding current which carries food particles to the mouth. Feeding and excretory currents have not been studied in detail. Associated with the mouth is a ciliated flap of tissue known as the epistome.

The digestive tract consists of an esophagus, stomach, intestine, and rectum. In some species, there is a distinct valve between the esophagus and stomach. The junction of the stomach and intestine occurs at the proximal or aboral extremity of the animal. The food seems to consist chiefly of microscopic phytoplankton. Diatom shells may frequently be found in the digestive tract and in fecal pellets.

There is a blood vascular system in which elliptical, nucleated corpuscles containing hemoglobin circulate. The vascular system consists basically of two longitudinal vessels, known as the afferent and efferent vessels, which are continuous with one another at the proximal end of the body. Distally, both vessels connect with a pair of semicircular vessels located at the level of the lophophore, immediately below the tentacles. Within each tentacle is a single, blind vessel which branches into two at its base and so connects with both semicircular

vessels. In the living animal, corpuscles can be seen pulsating up and down in the tentacular vessels. Associated with the longitudinal blood vessels is a blood sinus surrounding the gut and a large number of blind blood caeca which are particularly numerous in, or may be restricted to, the proximal end of the body. Associated with these caeca is the jellylike fat body, or vasoperitoneal tissue. Found among the large, semifluid cells of this tissue are inclusions of various sorts. Some of these probably consist of guanine or some related form of nitrogenous waste. Others may represent the products of hemoglobin breakdown.

The body cavity is subdivided by a series of longitudinal mesenteries extending from the digestive tract to the body wall. In most phoronids there are four such mesenteries occupying oral, anal, right lateral, and left lateral positions, thus dividing the coelomic cavity into four chambers. There is also a horizontal mesentery near the tentacular end of the body, separating a lophophoral coelom from the four larger and more proximal coelomic cavities.

The two nephridia open on either side of the anal papilla. Each nephridium consists of a duct, coiled once on itself and, usually, a pair of funnels, one opening into each of the oral and anal coelomic cavities. The funnels have extensive folded and ciliated margins.

The body wall consists of an outer layer of epithelial cells, many of them secretory and concerned with the building of the tube, and two layers of muscle. The outermost layer of muscle consists of circular fibers and inside this is a series of bundles of longitudinal fibers. An unpolarized nerve net underlies the external epithelium and continues as a more dense concentration in the form of a ring in the horizontal mesentery. Extending proximally from cell bodies in this ring are one or two giant nerve fibers which taper and disappear at the proximal end of the body. The giant nerve fibers are known in all species except one (*Phoronis ovalis*)

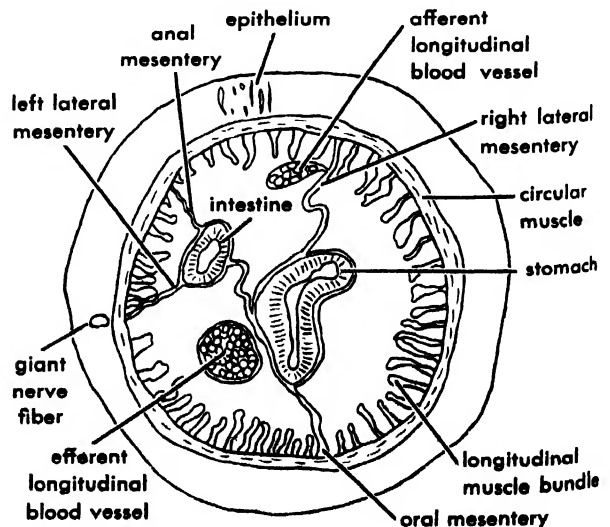


Fig. 2. Cross section through a *Phoronopsis*.

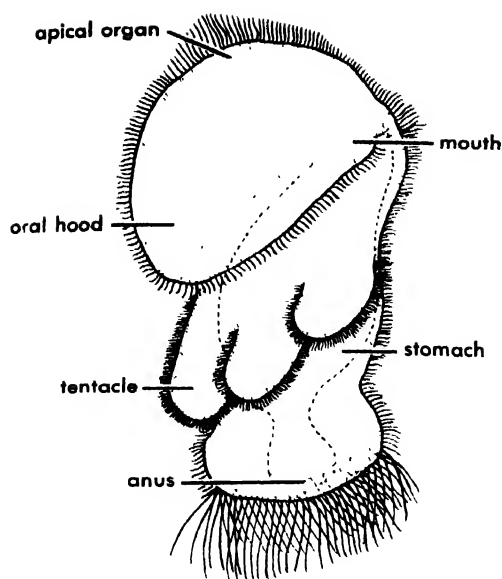


Fig. 3. Actinotroch larva of *Phoronis vancouverensis*.

and are probably concerned with the rapid retraction of the body into the tube.

**Reproduction.** Reproductive tissue is formed from cells which multiply first on the thin walls of the blood caeca. The phylum includes both dioecious animals and hermaphrodites. As the gonad increases in extent it displaces the vasoperitoneal tissue which shrinks proportionately. When ripe, the gametes are shed into the body cavity and find their way to the nephridia to pass through these organs to the outside. In at least one species (*Phoronis hippocrepia*), the ova are retained in the tentacular crown until the larval stage of development is reached. All phoronids may reproduce sexually, and in most cases the life history includes the pelagic actinotroch larva. Some species are known to reproduce asexually by transverse fission. See ANIMAL KINGDOM. [J.R.M.]

**Bibliography:** L. A. Borradaile, F. A. Potts, L. E. S. Eastham, and J. T. Saunders, *The Invertebrata*, 2d ed., 1935; S. F. Light, R. I. Smith, F. A. Pitelka, D. P. Abbott, and F. M. Weesner, *Intertidal Invertebrates of the Central California Coast*, 2d ed., 1954, G. E. MacGinitie, and N. MacGinitie, *Natural History of Marine Animals*, 1949.

## Phosphate

A negative ion having the formula  $\text{PO}_4^{3-}$ . Phosphates are derived from phosphoric acid,  $\text{H}_3\text{PO}_4$ .

The term phosphate is a broad term which encompasses all anions derived from acids containing phosphorus in the 5+ oxidation state as indicated in the list. All of those listed are obtained from  $\text{P}_4\text{O}_{10}$  and water.

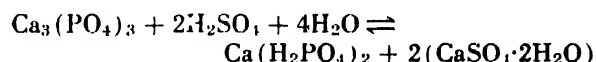
$(\text{HPO}_3)_n$	Metaphosphoric acid
$\text{H}_5\text{P}_3\text{O}_{10}$	Triphosphoric or tripolyphosphoric acid
$\text{H}_4\text{P}_2\text{O}_7$	Pyrophosphoric acid
$\text{H}_3\text{PO}_4$	Orthophosphoric acid

The naming of these phosphates is complicated by the fact that the acids contain several hydrogens which can be replaced stepwise by reaction with a base, and the fact that phosphates exist as polymers of the simpler acids listed. In the case of the most common acid, orthophosphoric, the salts are named as follows; phosphate means orthophosphate.

$\text{NaH}_2\text{PO}_4$	Monosodium phosphate Sodium dihydrogen phosphate Primary sodium phosphate
$\text{Na}_2\text{HPO}_4$	Disodium phosphate Sodium monohydrogen phosphate Secondary sodium phosphate
$\text{Na}_3\text{PO}_4$	Trisodium phosphate Tertiary sodium phosphate Normal sodium phosphate

The alkali metal phosphates and the primary alkaline-earth metal phosphates are soluble in water, whereas most other metal phosphates are practically insoluble at neutral pH.

A solution of trisodium phosphate is strongly basic and is used as a cleaning compound and water softener. Phosphates are important ingredients in commercial fertilizers. Natural phosphate rock can be converted into a useful fertilizer, superphosphate, by a reaction with sulfuric acid.



An important use of polymeric phosphates is as an ingredient in synthetic detergents and as sequestering agents.

The phosphate ion gives a yellow ammonium phosphomolybdate precipitate and yellow  $\text{Ag}_3\text{PO}_4$  precipitate which serve as analytical tests.

Certain organic phosphates have been used as insecticides and nerve gases. See FERTILIZER; ORGANOPHOSPHORUS COMPOUND; PHOSPHORUS.

[E.E.WR.]

## Phosphate metabolism

The reactive phosphates occur in the soft tissues of the animal body. The phosphates in mineralized tissues serve as an important storage depot, containing 75–85% of all the phosphorus in the animal body. The release of these stored phosphates in response to a lowered content in the blood plasma is not particularly effective. Hence, the level of inorganic phosphates in the blood plasma is relatively easily lowered when the dietary intake is inadequate, followed by the appearance of symptoms of aphosphorosis.

The central role of phosphates in life processes is indicated by their occurrence in ribo- and deoxyribonucleic acids, which are so important in protein synthesis and in the function of chromosomes in the processes of growth and heredity. The major significance of phosphates in metabolism is their role in the conservation and transfer of energy, particularly the energy produced in the tricarboxylic acid cycle (Krebs cycle) and in glycolysis.

They do so by participating in many phosphorylation and transphosphorylation reactions involving sugars and other organic compounds. See CHROMOSOME; KREBS CYCLE; NUCLEIC ACID.

In phosphorylation reactions, compounds are formed which yield relatively large amounts of free energy when their phosphate bonds are cleaved by hydrolysis. Examples of such compounds are creatine phosphate (CP) and adenosine triphosphate (ATP). A central part in the energy storage and transfer in all kinds of living tissue is played by ATP. In both CP and ATP, the phosphate bond can be transferred between molecules without liberation of inorganic phosphate. See ADENOSINETRIPHOSPHATE (ATP).

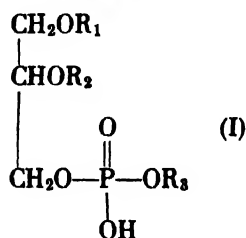
Phosphorus-containing coenzyme systems include the pyridine nucleotide system concerned with oxidation-reduction reactions; coenzyme A, the functional form of pantothenic acid, concerned, among other things, with transacetylation in conjunction with ATP and acetic acid; and the diphosphothiamine system, concerned with decarboxylation. See ACETYLCHOLINE; BIOCHEMISTRY; COENZYME. [H.H.M.]

**Bibliography:** W. D. McElroy and B. Glass (eds.), *A Symposium on Phosphorus Metabolism*, 1951.

## Phosphatide

A complex lipid containing phosphorus and in many cases nitrogen. Phosphatides are also known as phospholipids. The phosphatides are usually divided into groups on the basis of the nonlipid portion of the compound from which they are derived. For example, glycerophosphatides are derived from glycerophosphoric acid, sphingomyelins or phosphosphingosides are derived from sphingosinephosphate, and inositol lipids or inositol phosphatides are derived from inositol monoordiphosphate.

**Glycerophosphatides.** These are phosphatides which contain a glycerophosphoric acid residue. They are derived from glycerophosphoric acid (I) where  $R_1 = R_2 = R_3 = H$ . The following com-



pounds are glycerophosphatides: (1) phosphatidyl ethanolamine or cephalin where  $R_1 = R_2 =$

fatty acid,  $R_3 =$  ethanolamine; (2) phosphatidyl choline or lecithin where  $R_1 = R_2 =$  fatty acid,  $R_3 =$  choline; (3) phosphatidyl serine where  $R_1 = R_2 =$  fatty acid,  $R_3 =$  serine; (4) phosphatidyl inositol where  $R_1 = R_2 =$  fatty acid,  $R_3 =$  inositol; (5) lysophosphatidyl ethanolamine where  $R_1$  or  $R_2 =$  fatty acid,  $R_1$  or  $R_2 = H$ ,  $R_3 =$  ethanolamine; (6) lysophosphatidyl choline where  $R_1$  or  $R_2 =$  fatty acid,  $R_1$  or  $R_2 = H$ ,  $R_3 =$  choline; (7) plasmalogens where  $R_1$  or  $R_2 =$  fatty acid,  $R_1$  or  $R_2 = \alpha, \beta$ -unsaturated ether,  $R_3 =$  ethanolamine or serine; (8) ether lipid where  $R_1 =$  saturated ether,  $R_2 =$  fatty acid,  $R_3 =$  ethanolamine; (9) phosphatidic acid where  $R_1 = R_2 =$  fatty acid,  $R_3 = H$ ; (10) cardiolipin is a polymer of phosphatidic acid.

**Sphingophosphatides.** These are phosphatides which contain a sphingosine phosphate residue. There are two known members of this class—sphingomyelin (II) and phytoglycolipid which is considered in the section on glycolipids.

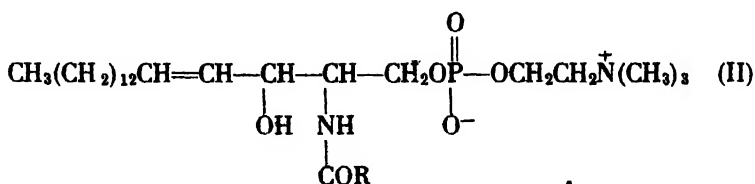
**Occurrence and functions.** Cephalin, lecithin, phosphatidyl inositol and the plasmalogens are present in both plant and animal tissues; phosphatidic acids and phytoglycolipid have been found only in plants; sphingomyelin has been found only in animal tissues; and the ether lipid was isolated recently from egg yolk.

Since an individual phosphatide may contain a variety of fatty acid residues it may be described as pure only with that limitation in mind. Phosphatides can act as protective colloids, wetting and emulsifying agents, and as antioxidants, and are therefore used considerably in the food and petroleum industries. The chief source of commercial phosphatides is soy bean. See LIPID. [H.E.C.; R.H.C.]

**Bibliography:** H. Wittcoff, *The Phosphatides*, 1951.

## Phosphorescence

Sometimes called afterglow, phosphorescence commonly denotes a delayed luminescence, that is, a luminescence that persists after removal of the exciting source. This original definition is rather imprecise, because the nature of the detector used will determine whether or not there is observable persistence. In a more rigorous sense, phosphorescence may be defined as delayed luminescence whose persistence time decreases with increasing temperature. In nonphotoconductive systems, phosphorescence arises when some process has placed the luminescent atom (or ion or molecule) in a metastable energy state (from which transitions to the state of lowest energy, or ground state, are forbidden) and energy has been provided to raise the









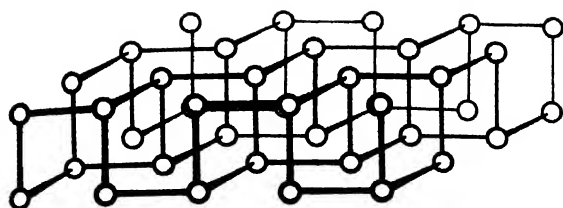


Fig. 5. Black phosphorus,  $P_n$ . (From R. E. Kirk and D. F. Othmer, eds., *Encyclopedia of Chemical Technology*, vol. 10, Interscience, 1953)

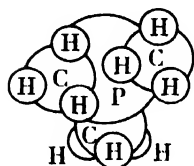


Fig. 6. Trimethyl phosphite,  $P(OCH_3)_3$ .

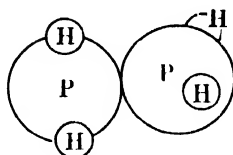


Fig. 7. Diphosphine,  $P_2H_4$ . (From R. E. Kirk and D. F. Othmer, eds., *Encyclopedia of Chemical Technology*, vol. 10, Interscience, 1953)

connected phosphorus and those based on triply connected phosphorus, there are also a few compounds in which there are 5 or 6 neighboring atoms bonded to the phosphorus. These compounds are very reactive and tend to be unstable because of the use of  $d$  orbitals in their  $\sigma$ -bond electronic structure. Examples are given in Figs. 8 and 9.

Structural reorganization plays an important role in the chemistry of phosphorus compounds. Thus, for example, when various mixtures of  $POBr_3$  and  $POCl_3$  are sealed in a glass tube and allowed to come to equilibrium, the intermediate compounds,  $POClBr_2$  and  $POCl_2Br$  are formed in various amounts depending on the ratio of the starting materials (Fig. 10). The  $POBr_3$ — $POCl_3$  reorganization involves compounds based on a single phosphorus atom to which is bonded 1 oxygen and 3 halogen atoms (chlorine and bromine are halogens). Structural reorganization also occurs between various members of a homologous series of compounds. In the polyphosphoryl chloride homologous series, reorganization takes place by exchange of bridging oxygen atoms with chlorine atoms, just as in the  $POBr_3$ — $POCl_3$  system the exchange is between chlorine and bromine atoms. The various structural units in a polyphosphoryl chloride composition are the monophosphorus compound,  $POCl_3$ ; the end group,  $Cl(O)PO_{\frac{1}{2}}$ ; the middle group,  $—O_{\frac{1}{2}}(Cl)P(O)O_{\frac{1}{2}}—$ ; and the branching group,  $OP(O_{\frac{1}{2}}—)_3$ , in which the bridging oxygen atoms are shown as  $O_{\frac{1}{2}}$ , since they are

shared between neighboring phosphorus atoms. A typical structure in a polyphosphoryl chloride is shown in Fig. 11.

When various ratios of chlorine to oxygen are employed, the distribution of the structural units changes as shown in Fig. 10, where  $A$  stands for the monophosphorus compound,  $POCl_3$ ;  $B$  for the ends;  $C$  for the middles;  $D$  for the branches, and  $D'$  for the completely branched compound, phosphorus pentoxide. The ends, middles, and branches do not exist by themselves, but must be combined together to form chemical compounds. The line labeled  $x$  in the figure represents the limit beyond which there is a sufficiently large proportion of branching points that infinite wall-to-wall molecular structures become statistically probable. The presence of such wall-to-wall molecular structures in the mixture of various sized and shaped polyphosphoryl chloride molecules leads to high viscosities and noticeable elastic behavior.

In spite of the fact that homologous series and compounds based on a number of phosphorus atoms are emphasized in this article, the extensive chemical literature before 1950 dealing with phosphorus chemistry was restricted almost entirely to compounds thought to be based on a single phosphorus atom (monophosphorus compounds).

**Principal compounds and uses.** Essentially all of the phosphorus used in commerce is in the form of phosphates. The majority of phosphatic fertilizers consist of highly impure monocalcium or dicalcium orthophosphate,  $Ca(H_2PO_4)_2$  and  $CaHPO_4$ . These phosphates are salts of orthophosphoric acid, which is the monophosphorus compound in the phosphate homologous series. Impure dicalcium orthophosphate for fertilizer use is usually called superphosphate, whereas the impure monocalcium phosphate used in this

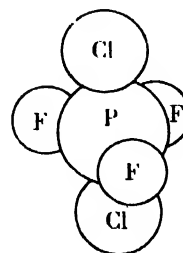


Fig. 8. Phosphorus dichloride trifluoride,  $PCl_2F_3$ . (From R. E. Kirk and D. F. Othmer, eds., *Encyclopedia of Chemical Technology*, vol. 10, Interscience, 1953)

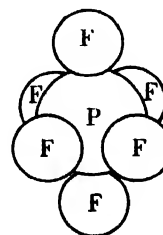


Fig. 9. Hexafluorophosphate anion,  $PF_6^-$ .

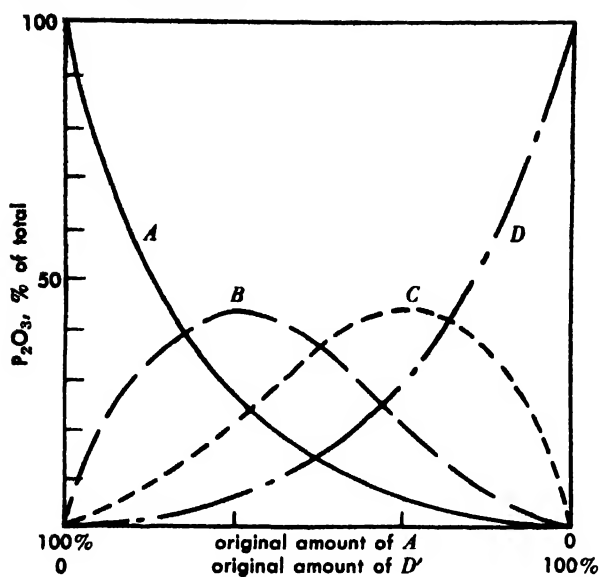


Fig. 10. Reorganization equilibria where *A* is  $\text{POCl}_3$ , *B* is  $\text{POCl}_2\text{Br}$ , *C* is  $\text{POClBr}_2$ , and *D* and *D'* are  $\text{POBr}_3$ .

application is called triple superphosphate. See FERTILIZER.

Two properties of the family of chain phosphates have led to numerous industrial applications for these compounds. These properties are deflocculation of colloidal particles and formation of soluble complexes with cations. The chain phosphates are strongly adsorbed on the surfaces of inorganic solids, and hence, give these surfaces high negative charges. When finely divided particles bear such high charges, they repel each other and are deflocculated, peptized, or dispersed. An interesting example of this phenomenon is found when a plastic clay-water mass is treated with a chain phosphate. By addition of, perhaps, a few tenths of 1% of sodium tripolyphosphate to a plastic mass of clay suitably rigid for sculpturing, the clay particles are deflocculated so that the mass liquefies to a consistency similar to that of tomato soup.

The formation of soluble complexes with cations has often been described under the term sequestration, because a complexed ion is sequestered or hidden away in the solution so that it no longer exhibits its normal chemical reactions. The calcium and magnesium of hard water are sequestered by the addition of small (stoichiometric) amounts of chain phosphates so that the water is effectively softened. The complexed calcium will then no longer form precipitates with the carbonate or sulfate in the water to give pipe scale, or with soap anions to give, for example, a ring around the bathtub.

The third member of the family of sodium phosphates, sodium tripolyphosphate, is the major compound used in building synthetic detergents to achieve improved cleaning, primarily by dispersing inorganic soil and softening the water. The average household detergent produced in the United States

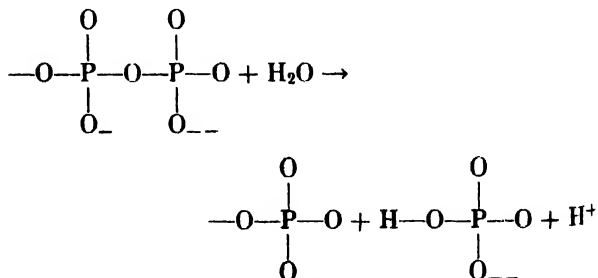
for washing clothes consists of 50% by weight of sodium tripolyphosphate,  $\text{Na}_5\text{P}_3\text{O}_{10}$ . This compound is used extensively in water softening, as are other members of the homologous series of chain phosphates. See DETERGENT; SURFACE-ACTIVE AGENT; WATER SOFTENING.

An interesting water-softening application is found in "threshold treatment" in which tiny traces of a chain phosphate (much less than would be used in sequestering) are used to prevent the formation of pipe scale from hard waters. This application is related to the dispersing action of the phosphates, because traces of phosphate adsorb on the growing surface of the pipe scale as it begins to form, and this inhibits its further growth.

A major pharmaceutical use of phosphates is in toothpastes, in which dicalcium phosphate is the most popular polishing agent. Monocalcium phosphate and sodium acid pyrophosphate,  $\text{Na}_2\text{H}_2\text{P}_2\text{O}_7$  (the pyrophosphate is the second member of the phosphate family), are employed as leavening agents in cake mixes, refrigerated biscuits, self-rising flour, and baking powder.

Special mixtures based on orthophosphoric acid,  $\text{H}_3\text{PO}_4$ , are used to phosphatize metal surfaces. In this treatment, the surfaces become covered with a thin adhering layer of insoluble orthophosphate salts which protect the metal from corrosion and offer an especially adherent base for painting. Automobile bodies, for example, are now generally phosphatized before they are painted, to prevent rusting in use. Orthophosphate esters find wide use as plasticizers having flame-proofing properties, and as gasoline and oil additives.

The phosphorus compound of major biological importance is adenosine triphosphate, which is an ester of the salt, sodium tripolyphosphate, widely employed in detergents and water-softening compounds. Practically every reaction in metabolism and photosynthesis involves the hydrolysis of this tripolyphosphate to its pyrophosphate derivative, called adenosine diphosphate. The hydrolysis of chain phosphates occurs through splitting of a  $\text{P}-\text{O}-\text{P}$  linkage as indicated in the following chemical equation:



In neutral solution at room temperature, the rate for this process is extremely slow. However, enzymes increase the rate many thousandfold. The equilibrium between adenosine triphosphate, water, adenosine diphosphate, and the orthophosphate ion is strongly shifted toward the hydrolysis product,

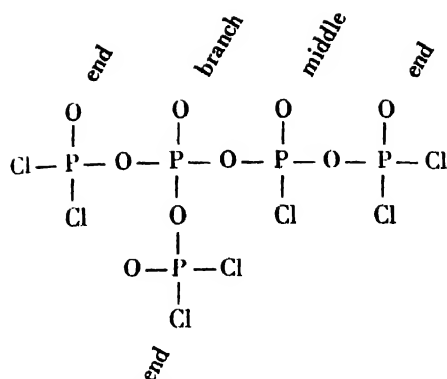


Fig. 11. Isopentapolyphosphoryl chloride,  $P_5O_8Cl_7$ .

adenosine diphosphate and the orthophosphate ion. Because of these facts, organic reactions in biological systems are naturally controlled so that life can exist. See ADENOSINETRIPHOSPHATE (ATP).

[J.V.W.]

*Bibliography:* J. R. Van Wazer, *Phosphorus and Its Compounds*, vol. 1, 1958.

## Phot

The unit of illumination when the square centimeter is taken as the unit of area. It is equal to the density of one lumen per square centimeter and is defined in a similar manner as the foot-candle and the lux. See FOOT-CANDLE; ILLUMINATION; LUX.

[R.C.P.]

## Photochemistry

The branch of chemistry dealing with the interrelationships between light and chemical reactions. Photochemistry includes the study of chemical reactions producing, or produced by, visible and near ultraviolet light of wavelengths between those of infrared light and x-rays. This region of the spectrum includes sunlight or solar energy as it reaches the earth's crust, namely, the near infrared, visible, and near ultraviolet part of sunlight between 12,000 and 3000 angstroms (Å), and sunlight as it exists at reasonable intensity in outer space, namely, down to about 1000 Å. The maximum intensity of sunlight is in the red part of the spectrum between 6000 and 12,000 Å. Radiation chemistry pertains to studies of the chemical reactions produced by x-rays, γ-rays, and particles such as electrons of about the same or higher energy equivalent.

Common photochemical reactions are the natural photosynthetic process and photography. Chemical reactions producing light are usually identified as burning, or combustion. The chemical reaction in the firefly and other chemiluminescent reactions, however, produce cold light.

In the natural photosynthetic process the reactions are brought about by light absorbed by chlorophyll. The over-all reaction should be written as



in order to indicate that all the oxygen gas comes from the water and none from the carbon dioxide. This has been proved by employing the oxygen isotope of mass 18 to follow the path of oxygen in the process. The path of carbon in the process has been followed by employing the radioactive carbon isotope of mass 14. Very little is known, however, about the way in which the light absorbed by the chlorophyll brings about the reactions.

The unit of light energy most useful in photochemistry is the photon,  $\epsilon = hc/\lambda$ , where  $h$  is Planck's constant ( $6.5 \times 10^{-27}$  erg sec),  $c$  is the velocity of light ( $3 \times 10^{10}$  cm/sec) and  $\lambda$  is the wavelength of the light in cm ( $1\text{Å} = 10^{-8}$  cm).

Another photochemical unit of light energy is the einstein which is the energy of  $6 \times 10^{23}$ , or 1 mole  $N$ , of light quanta. Thus 1 einstein of red light is  $Ne = Nhc/\lambda = (6 \times 10^{23} \times 6.5 \times 10^{-27} \times 3 \times 10^{10}) / (6700 \text{ Å} \times 10^{-8}) = 17.5 \times 10^{11}$  ergs or  $(17.5 \times 10^{11}) / (4.186 \times 10^7) = 42,000$  cal. This amount of energy is greater than the activation energy required to initiate many thermal reactions.

**Quantum yield.** The efficiency of a photochemical reaction is usually expressed in terms of the quantum yield, which is equal to the number of moles of the stated reactant disappearing, or to the number of moles of the stated product produced, per einstein of light of the stated wavelength absorbed. The gross quantum yield is calculated from the light absorbed by the entire photosensitive system. Net quantum yield is based on the light absorbed by the stated component or species of the system. Photochemical reactions are most easily understood in terms of net quantum yields.

**Photochemical reactions.** Photochemistry may also be defined as the chemistry of energy-rich, photon-excited states. When produced by the absorption of visible or ultraviolet light, these are electronically excited states resulting from the transfer of an electron to a higher energy level. The time required for this transfer is so short that during this time the positions of the nuclei of the atoms involved remain unchanged. This is known as the Franck-Condon principle.

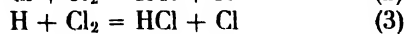
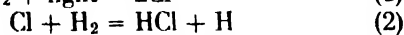
Following this primary act, one of a great many different kinds of processes may take place:

1. Production of energy-rich singlet and triplet states, followed by fluorescence, phosphorescence, and degradation of the absorbed energy to heat.
2. Transfer of the electronic energy, or transfer of electrons, photons, or hydrogen atoms, to or from other species.
3. Breaking apart of excited species such as  $Cl_2$  into Cl atoms,  $O_3$  into  $O_2$  and O.

4. Production of a wide variety of chemical reactions. Among these reactions are the electron transfer reaction in which water is decomposed into hydrogen and oxygen by light absorbed by cerium ions in water; cis to trans, and trans to cis, isomerization as in the case of thiindigo; shifting of the positions of double bonds as when ergosterol is converted into vitamin D; dissociation of the excited

species into atoms, ions, molecules, and radicals; changes in the acidic or basic strength of the light-absorbing species; polymerization reactions; germicidal action; sunburn and tanning.

In its broadest sense photochemistry may properly be said to deal with the study of any of these processes. In a narrower sense, however, photochemistry deals only with the chemical reactions brought about by absorbed light. This includes studies of the kinetics and mechanisms of reactions such as that between  $H_2$  and  $Cl_2$  to produce  $HCl$ . In this case, only  $Cl_2$  absorbs visible and near ultraviolet light and the reaction proceeds mainly as follows:



The net quantum yield for this chain reaction is over 1,000,000 under favorable conditions.

**Light absorption.** The fraction of the light absorbed by a stated component or species in a system requires a knowledge of the relative concentrations and light-absorbing powers of all the species in the system which absorb a significant amount of light. Concentrations for this purpose are stated in terms of the number of light-absorbing species per unit volume, for example, in moles per liter. Light-absorbing powers are expressed in terms of a constant which is characteristic of the stated species in the stated environment at the stated wavelength such as the molar absorptivity,  $\epsilon_i = A_i/c_i$ , where  $A_i = \log_{10} I_0/I$ ,  $c_i$  is the concentration of the stated species in the stated environment, and  $l$  is the length of the light path over which the light intensity of the stated wavelength decreases from an initial intensity  $I_0$  to  $I$  because light is absorbed by the stated species. See SPECTROPHOTOMETRIC ANALYSIS.

**Light sources.** The interpretation of a photochemical reaction is greatly simplified when concentrations are uniform throughout the reacting mixture. This requires that the light fill the whole system and be weakly absorbed, or that there be adequate mixing, especially in the parts of the system absorbing most of the light.

Light intensities, light-absorbing powers, and quantum yields sometimes change rapidly as a function of wavelength, so that quantitative photochemical studies are best carried out with monochromatic light or with light consisting of a suitably small range of wavelengths.

Monochromatic light is conveniently obtained by employing atomic light sources which emit the desired wavelength as part of a discontinuous spectrum. The desired light must be sufficiently different in wavelength from the other emitted rays so that it can be isolated easily at a relatively high intensity. A common light source is the mercury arc lamp.

Monochromatic light intensities obtained from most light sources are usually low. Therefore, the success of a photochemical study often depends

upon the proper design of an apparatus for isolating and bringing to bear upon a sufficiently small volume of the photosensitive system, most of the light of the desired wavelength emitted from the light source. Monochromatic light is isolated successfully for photochemical studies by means of filters or monochromators and occasionally by means of focal isolation.

Whenever possible, advantage is taken of the fact that the photosensitive system may absorb a suitably small range of wavelengths of the light incident upon it, although the latter may consist of a very wide range of wavelengths. Under these conditions, however, the evaluation of the light absorbed by the system is especially difficult.

Measurement of the light absorbed by a system has been accomplished by means of chemical actinometers, bolometers, thermopiles, and phototubes with proper auxiliary equipment. The uranyl oxalate and ferric oxalate actinometers are convenient and reliable.

**Energy relationships.** It is sometimes convenient to think of a photochemical reaction in a liquid system as being initiated in a photochemical cluster not unlike the critical complex of thermal reactions. There is, however, one important difference namely, that the products of a photochemical reaction may contain as chemical energy a significant fraction of the energy of the absorbed light, even when the reaction is essentially complete, whereas thermal reactions do not take place to any significant extent if the free energy of the products is greater than the free energy of the reactants. The latter is also true of photochemical reactions when one includes the energy of the absorbed light as part of the free energy of the reactants.

The elucidation of photochemical reactions is often easier in terms of changes in net quantum yields than of changes in rate constants. Also it is often possible to identify the part of the light-absorbing species responsible for the light absorption and thereby obtain information about intramolecular as well as intermolecular energy transfer processes and accompanying thermal reactions. See FREE RADICAL; LIGHT; LUMINESCENCE; PHOTOGRAPHY; PHOTOSYNTHESIS; RADIATION CHEMISTRY. SPECTROSCOPY. [L.J.H.]

**Bibliography:** F. Basolo and R. G. Pearson, *Mechanisms of Inorganic Reactions*, 1958; E. J. Bowen, *The Chemical Aspects of Light*, 2d ed., 1946; L. J. Heidt, *Converting solar to chemical energy*, *Proceedings of the World Symposium on Applied Solar Energy*, 1956; C. Reid, *Excited States in Chemistry and Biology*, 1957; A. Weissberger (ed.), *Technique of Organic Chemistry*, vol. 2 2d ed., 1956.

## Photoclinometer

A term applied to directional surveying instruments which record photographically the direction and magnitude of well deviations from the vertical. Two instruments of this type are in wide use, the

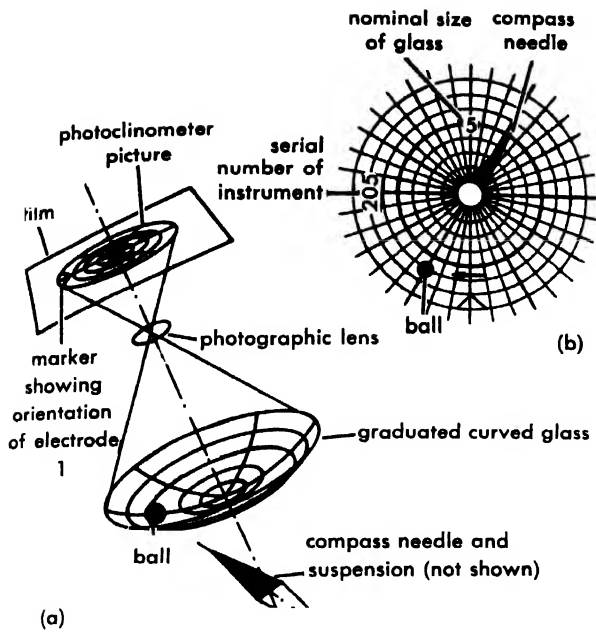


Fig. 1. (a) Principal features of the Schlumberger Photoclinometer. (b) Type of record obtained with the Schlumberger Photoclinometer. (Schlumberger Well Surveying Corp.)

Schlumberger Photoclinometer and the Surwell Clinograph. Both instruments record a series of deviation measurements on one trip into and out of the well. From this series of data it is possible to plot quite accurately the course of the well.

In the Schlumberger Photoclinometer the deviation from the vertical is indicated by a small metal ball which rolls in a transparent glass bowl graduated in circular degrees. The direction of the deviation in azimuth is indicated by a magnetic compass. With the instrument suspended by an electrical cable, the positions of the compass and steel ball are photographed on a 35-mm film by operation of electrical controls at the surface which turn on lights in the instrument and snap the camera shutter. After the picture is taken the film is moved to a new position. Pictures can be taken at a rate of about one per minute. Correlation of the pictures with the depths at which they are taken

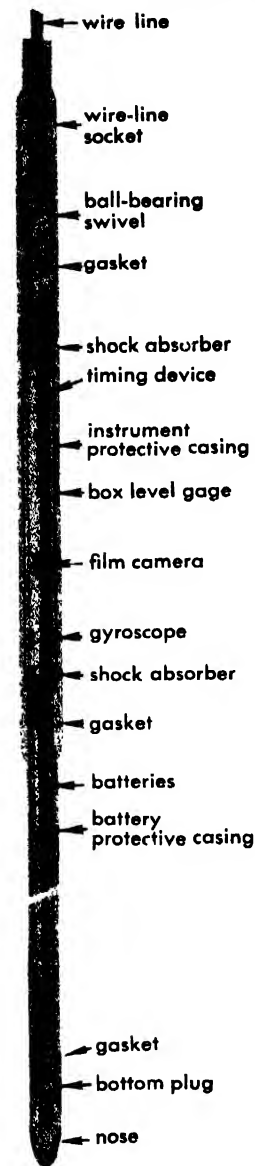


Fig. 3. Vertical section through Surwell clinograph. (Sperry-Sun Well Surveying Co.)

(known by the length of the suspending cable) yields a measure of the magnitude and direction of deviation of the hole as a function of depth.

Fig. 2. Motion-picture film records made by the Surwell clinograph. (Sperry-Sun Well Surveying Co.)

The Surwell Clinograph also operates electrically but is powered by batteries contained in the instrument. The deviation from the vertical is indicated by a box level gage and the direction in azimuth by a gyroscopic compass permitting its use inside steel pipe. This operation is not possible when a magnetic compass is used unless the pipe is made of special nonmagnetic steel. Since the instrument also contains a watch and a dial thermometer, a simultaneous record of amount and direction of deviation, temperature, and time can be made on 16-mm film. Readings are taken, both descending and ascending, at regular intervals which are preset on the instrument before it is lowered on a wire line into the well, thus providing a check on accuracy. Level gages having maximum inclinations of 20, 40, and 55° respectively are provided to be used according to the magnitude of deviation.

[H.G.BO.]

**Bibliography:** L. C. Uren, *Petroleum Production Engineering: Oil Field Development*, 4th ed., 1956.

### Photoconductive cell

A device for detecting or measuring electromagnetic radiation by variation of the conductivity of a substance (called a photoconductor) upon absorption of the radiation by this substance.

To detect or measure the radiation, the cell is connected in series with an electrical source and a galvanometer. The current through the cell is a function of the intensity of the radiation falling on the cell. The galvanometer measures the current.

Photoconductors can be classified into the elemental types, such as selenium, iodine, boron, diamond, germanium, and silicon; and the compound types, such as the sulfides, selenides, and tellurides of lead, thallium, and cadmium. Each of these materials must be doped with the proper amount of a selected impurity.

The photoconductor is usually prepared in the form of a thin film by evaporation of the material under vacuum, by chemical precipitation or by sintering of the powdered material.

Cadmium sulfide cells are extensively used in the visible spectrum for industrial applications because of their high sensitivity. However, they have a certain inertia, and their response depends upon previous exposure to light (hysteresis). They also generate a relatively high current when not illuminated (dark current).

The antimony sulfide cells are less sensitive than the cadmium sulfide cells, but their response time is much shorter and they do not display any hysteresis.

Lead sulfide and lead selenide cells are specially sensitive to infrared radiation, and this sensitivity toward the higher wavelengths increases at low temperature. For measurement in the far infrared, certain cells are cooled with liquid hydrogen (−250°C).

Photoconductive cells are characterized by their sensitivity in infrared (1–25 microns) and their

short response time. They are used for high-speed recording, high-resolution spectroscopy, television, electrophotography, and as infrared detectors. For more detailed information see INFRARED DETECTOR; see also PHOTOCONDUCTIVITY; PHOTOELECTRIC DEVICES.

[J.J.RO.]

**Bibliography:** R. G. Breckenbridge (ed.), *Photoconductivity Conference*, 1956.

### Photoconductivity

The increase in electrical conductivity displayed by many nonmetallic solids when they absorb electromagnetic radiation. The radiation may lie in any part of the spectrum from the infrared to the x-ray and γ-ray region. Photoconduction may proceed by several different mechanisms, depending on the type, the composition, and the crystal perfection of the solid involved. Photoconduction finds considerable practical application in television cameras, infrared detectors, light meters, and indirectly in the photographic process.

**Alkali halides.** Photoconductivity due to color centers in alkali halides (frequently called primary photoconduction) occurs in crystals such as common rock salt (sodium chloride) if they have been heated in sodium or other alkali metal vapor. This treatment gives rise to imperfections called color centers which color the crystal. These centers are lattice sites at which electrons take the place of missing negative ions; the color centers absorb visible light (see COLOR CENTERS). As a result, the electrons are set free, and they are set in motion when an electric field is applied to the crystal. This motion induces electric charges on the electrodes that supply the field. Current flows in the external circuit even though no charges pass from the electrodes into the crystal.

After being set free by the light, the electrons usually move only a short distance before they are stopped. This happens mainly at other color centers. The distance over which the electrons move is called the range, and it increases as the applied field is made stronger. The photoconduction is approximately proportional to the field strength as long as the electron range (in a typical case 10<sup>−5</sup> cm) is shorter than the sample length. When this is no longer true, the freed electrons move to the end of the sample. At this point, the photoconduction is constant with increasing applied field.

This photoconduction is excited most easily by photon energies lying in the optical absorption peak of the color centers. For potassium iodide, a typical alkali halide, this peak is centered near 1.6 ev, in the red region of the spectrum. Primary photoconductivity can also occur, however, at higher photon energies. Above 2.5 ev, in the blue and ultraviolet spectral regions, the electrons ejected from the color centers have enough energy to escape through the crystal surface. They then contribute to photoemission. See PHOTOEMISSION.

Primary photoconductivity usually occurs for only a relatively short time in an alkali halide. If



the current flows for too long (say 4 min), a much more complex phenomenon called secondary photoconductivity may result.

Exciton-induced photoconductivity occurs in alkali halides that contain color centers if the incident radiation lies in the first intrinsic (or fundamental) optical absorption peak of the crystal itself. (For potassium iodide, this peak lies at a photon energy of 5.6 eV in the far ultraviolet.) The effect is like exciton-induced photoemission, except that the excited electrons remain in the crystal. It occurs in two stages. The absorbed photons produce excitons, which are electrically neutral entities. These then transfer enough energy to color centers to eject electrons. Thereafter, the process is similar to primary photoconductivity. Exciton-induced photoconduction occurs also in crystals such as barium oxide. *See* EXCITON.

Intrinsic photoconductivity in alkali halides takes place when light is absorbed in the ideal pure crystal lattice with the resultant production of mobile electrons and positive holes (*see* HOLES IN SOLIDS). In the absence of defects such as color centers, photoconduction does not occur at energies corresponding to the first optical absorption peak of an alkali halide. The concept of excitons as neutral "particles" developed as it did because of this experimental fact. At a slightly higher photon energy, one expects intrinsic photoconductivity to set in. It is analogous to the intrinsic photoemission which begins in potassium iodide at about 7 eV. Photoconduction should set in at slightly lower photon energies than photoemission. It is difficult to measure accurately.

**Silver halides.** Photons having energies just high enough to produce optical absorption in a pure silver halide also produce electrons and holes. Intrinsic photoconductivity results. Thus, the silver halides stand in contrast to pure alkali halides, in which optical absorption in the first fundamental absorption band does not produce intrinsic photoconductivity. The motion of positive holes in silver chloride, a typical silver halide, is apparently not appreciable. The electrons, on the other hand, can move as much as 1 cm in experimentally feasible electric fields. The motion of electrons is often limited by electron traps (*see* TRAPS IN SOLIDS). The electron migration then resembles that for extrinsic photoconductivity caused by color centers in alkali halides. At photon energies near 6 eV, the excited electrons are so energetic that some of them escape through the surface, and intrinsic photoemission results.

Photoconduction with consequent ionic motion in the silver halides plays an important role in the photographic process. *See* PHOTOGRAPHY.

**Germanium.** Intrinsic excitation of photoconductivity in germanium, a typical elemental semiconductor, occurs for photon energies in the fundamental (intrinsic) optical absorption band. This band has an edge at about 0.7 eV in the infrared, and it extends continuously through higher photon

energies in the visible and ultraviolet regions. Electrodes are connected to the germanium crystal to furnish an electric field. Electrons or positive holes can pass from these connections into the crystal, and an electric current flows through the specimen in the dark. When the crystal is illuminated, additional electrons and holes are created in equal numbers. In general, both these excess-current carriers move in the electric field and contribute to the photoconductivity. An important consideration is that the sample must contain them in equal numbers (this is called the charge neutrality condition); otherwise, prohibitive electric fields would build up. Thus, if an electron leaves the sample through one electrode, another electron enters at the other end. Alternatively, an electron and hole may annihilate one another in a recombination process. Direct recombination is possible, but not very probable. Almost all of the recombination takes place at defects or impurities called recombination centers, some of which may be at the sample surface. If the illumination is turned off, the concentration of excess electrons and holes (and therefore the photoconductive current) disappears as a function of time in an exponential way. It decreases by a factor of  $e = 2.718 \dots$  in an interval  $t$ , which is called the lifetime for electron-hole pairs. This time is determined by the number and type of recombination centers in the particular sample. For very pure germanium crystals at ordinary temperatures, the pair lifetime is 1 msec or higher.

To calculate the magnitude of the photoconductive current, it is convenient to consider a cube of germanium 1 cm on a side with electrodes on two opposite faces. If  $N$  photons/sec are absorbed uniformly in this volume, the number of excess free electrons and holes reaches a steady concentration  $Nt$  (for each carrier type). The number of electrons  $P$  flowing per second through the battery in the external circuit is

$$P = V N t (\mu_e + \mu_h)$$

where  $V$  is the battery voltage, and  $\mu_e$  and  $\mu_h$  are the mobilities of the electrons and holes respectively. The ratio  $G = P/N$  determines the sensitivity, and is called the photoconductive gain factor. Now the transit time  $T_e$  required for an electron to traverse the germanium sample is  $1/V\mu_e$ ; that for holes is  $T_h = 1/V\mu_h$ . Thus the gain may be expressed as

$$G = t \left( \frac{1}{T_e} + \frac{1}{T_h} \right)$$

If the germanium cube is at ordinary temperature, with  $t = 10^{-3}$  sec and  $V = 1$  volt,  $G$  is about 5.

Photoconduction of this same general kind occurs in silicon and in certain compounds such as indium antimonide.

High-gain photoconduction can occur when electron or hole traps are present in a crystal of germanium. For example, nickel atoms deliberately

added as impurities behave as hole traps in germanium near the temperature of liquid nitrogen. If an appropriate amount of arsenic is also present, each Ni atom becomes a doubly negative ion,  $\text{Ni}^{2-}$ . Because of its strong negative electric charge, it repels electrons and attracts positive holes. Thus, when illumination sets both holes and electrons free in the sample, the  $\text{Ni}^{2-}$  ions quickly capture holes. Accordingly, the doubly negative  $\text{Ni}^{2-}$  becomes singly negative  $\text{Ni}^{-}$ . It is still negative and it still repels electrons. Thus, recombination of electrons with the captured holes is drastically reduced, and the hole is said to be trapped. It is immobile and does not contribute to photoconduction. However, for each trapped hole, a mobile electron is held in the crystal to preserve the condition of charge neutrality. Recombination occurs after a comparatively long time, called the free-electron lifetime,  $\tau_e$ . It determines the speed with which the photoconductor responds to changes in illumination. The photoconductive gain is in this case  $t_e/T_e$ , where  $T_e$ , as before, is the electron transit time. At the low temperature considered here,  $T_e$  is about  $10^{-4}$  sec if the sample is a cube of unit volume and if 1 volt is applied by the battery. Thus, the gain becomes  $10^4$ , and for every photon absorbed in the sample, a great many electrons flow through the external circuit. The gain increases as the applied voltage increases, but complications prevent it from increasing indefinitely.

High-gain photoconductivity of the same general character occurs in many other materials, such as cadmium sulfide, cadmium selenide, and lead sulfide. It was recognized in many of these before being studied in germanium. It is not understood quite as precisely as in germanium because the traps are not yet as well identified and because it is more difficult to control the composition and perfection of the crystals. On the other hand, high gain can occur in cadmium sulfide, for instance, at ordinary temperatures, and is important for applications in photoconductive devices. For the germanium photoconductors, a high sensitivity or gain corresponds to a proportionately long response time. In more complex photoconductors, the response may be sluggish even when the gain is low. This usually means that both electrons and holes are being trapped, frequently in complex fashion.

In germanium and other semiconductors, extrinsic excitation of photoconductivity occurs when radiation ejects electrons or holes directly from impurities into the conduction band or the valence band. Photon energies may be much less than those in the intrinsic optical absorption range, and the photoconductor may respond much farther into the infrared.

In certain cases, the condition of charge neutrality may not be satisfied in semiconductors. The photoconductivity then behaves in a more complicated manner than outlined here. See ABSORPTION (ELECTROMAGNETIC RADIATION); LUMINESCENCE. [L.A.]

**Bibliography:** R. G. Breckenridge, B. R. Russell and E. E. Hahn (eds.), *Photoconductivity Confer-*

*ence*, 1956; C. Kittel, *Introduction to Solid State Physics*, 2d ed., 1956; F. Seitz and D. Turnbull (eds.), *Solid State Physics*, vol. 8, 1959.

## Photocopying processes

Those means by which a copy is created on a sensitive surface (generally paper, film, or metal plate) by the action of light. The term is generally applied only to documentary reproduction. It is distinct from the photographing of gross objects (portraiture, for example), from cinematography, and from other highly specialized applications of photography, although these applications frequently overlap or are combined with photocopying of documents. The document to be photocopied must already have been prepared by other applications of photography, by manuscript, or by typewriter. A document in this case is classed as either a line drawing or continuous tone illustration, or a combination. Some photocopying processes do not handle tone satisfactorily.

Photocopying offers practical printing methods for the production of a single copy or a limited number of copies, or for the production of a stencil or master from which to run off larger numbers by use of diazo paper or offset lithography. It falls into several chemically distinct processes, distinguished by the chemistry of the photographic material and its development and fixation. Several sensitive materials are used: silver halide salts, diazonium salts, and ferric salt; newer methods of exposure apply infrared radiation, electrostatics, and electrolysis. The last three forms, along with the diazo when developed by exposure to ammonia gas, are referred to as dry processes, as opposed to the more common wet processes which use liquids. The basic camera techniques of photocopying are not so distinctive as is the mechanical equipment developed for use with the many photographic materials and the vast variety of their chemical characteristics.

For a discussion of the development of photography and a full description of specific photographic materials see PHOTOGRAPHIC MATERIALS; PHOTOGRAPHY.

Photocopying is now applied in business offices and libraries, in diverse problems of industrial production, as well as in sophisticated data-processing systems. Advantages of photocopying are its photographic accuracy except for occasional problems with color; reduction and enlargement ability of some processes; speed in most instances; space saving in the case of microfilm; economy of labor and materials over any other short-run copying process; convenience of handling the thin flexible material as opposed to letterpress type or electrolytic plates; simplicity of machine operation by untrained staff for certain processes; and flexibility as achieved by the combination of photocopying processes with other printing processes.

Photocopying processes may be somewhat arbitrarily divided into seven classes: silver halide photocopying, transfer processes, thermography (tech-

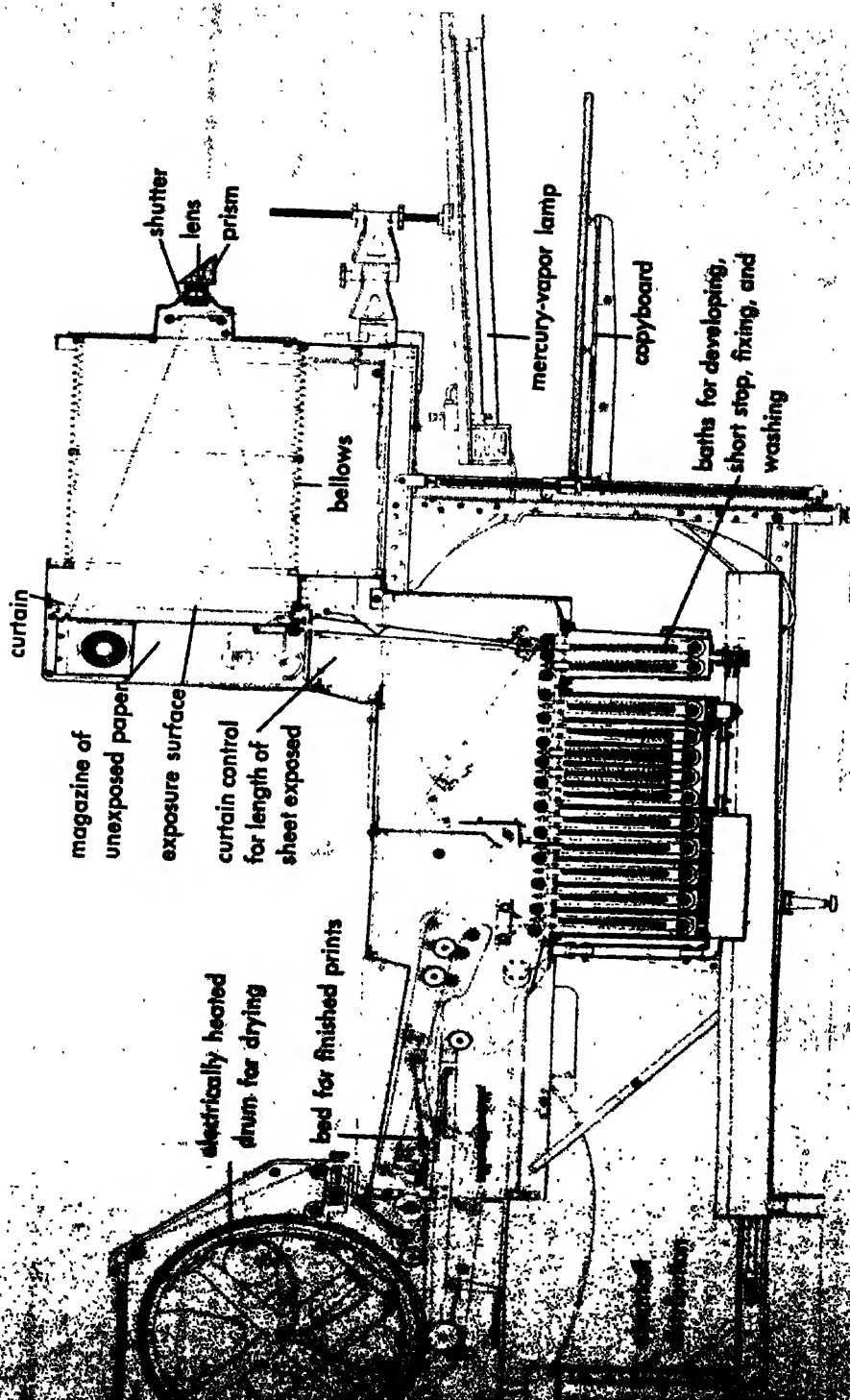


Fig. 1. Diagram of a continuous prismatic photocopier. (Photostat Corp.)

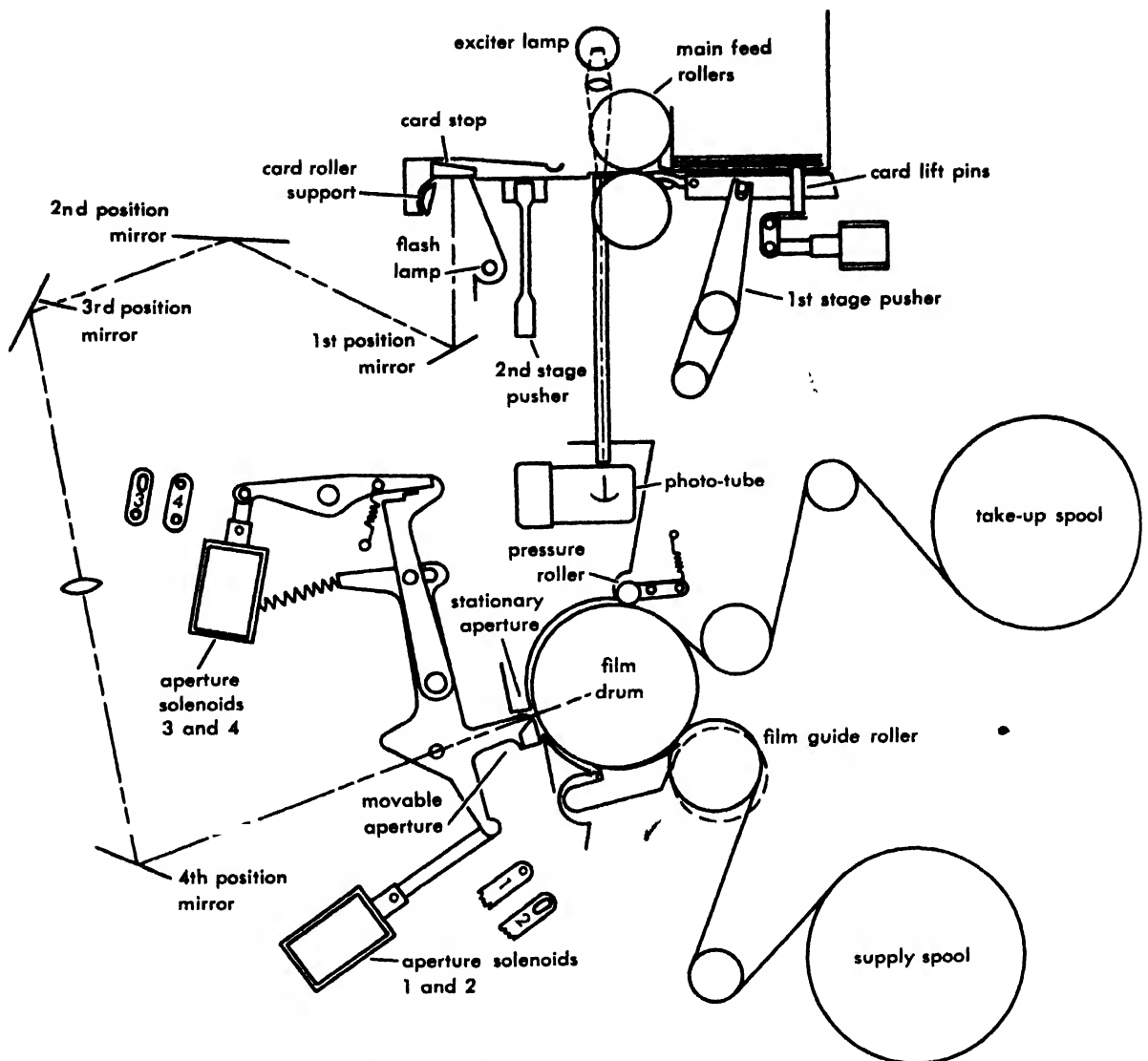


Fig. 2. Diagram of Listomatic camera film aperture control. (Recordak Corp.)

nically not photographic but a similar technique), plan copying, electrostatic processes, the electrolytic process, and microfilming.

The seven classes are not mutually exclusive because one may be used in combination with another. This is frequently the case with microfilming, in which the transparency may be the intermediate for enlargement to one of a variety of end products, for example, bound volumes printed by continuous xerography, offset plates, and multilith master. Transfer processes and thermography are referred to as rapid copy or office-type copying because they are relatively or completely dry, require no darkened laboratory, and are mechanically simple for an office assistant to operate. Facsimile reproduction is sometimes erroneously classed as photocopying when it is actually a mechanical or chemical transcription of electrical impulses generated by a photoelectric-cell scanning device; yet experiments are now combining this communications method with photocopying as the end process.

**Silver halide photocopying.** This is the familiar technique which has been developed into various forms, including microfilm. The simplest is contact exposure of an original with negative or direct positive paper in front of a light box, such as the photoprinting machines designed for making blueprints or whiteprints. Add a nonadjustable bellows and adjustable lens, and it approximates the hand camera, such as the Leica which has been commonly used to copy documents since 1924. With an adjustable bellows, it resembles the process camera which is so admirably adapted to photomechanical graphic reproduction demanding enlargement, reduction, and excellent handling of tone.

The next stage of sophistication uses a prismatic mirror to obtain a direct-reading negative instead of the usual inverted negative. Since its introduction in 1906, this has been the most frequently used method of photocopying. The best-known machines are the Dexigraph, Photoclerk, Photostat, Rectigraph, Rutherstat, and Statmaster (Fig. 1). Some

of these have fixed focus or are capable of only slight reductions; some automatically process, cut, and dry the prints. The most elaborate equipment can be used to make copy negatives and prints, slides, halftone prints from color transparencies, screened Veloxes, and stripped-in prints for photo-offset copy (see PRINTING PLATE). In some models, use of direct-reading positive paper is possible.

Photocomposing is a process which serves as a typesetting substitute. Characters on a transparent surface are brought to a correct position over the copy being made. In some devices, this positioning is accomplished by means of a punched tape control which indicates letters, face, point size, and justification. A strobe flash projects the image through appropriate lenses onto the sensitive silver film or paper; and from the developed film, plates are prepared, usually for gravure or photolithography. Since such a process was first patented in 1877, more than 50 variant schemes have been proposed. Most of the machines now available use the standard typewriter layout for initial composition, and most have their own correction units. This process has been used for credit lines of motion picture film. Use for newspaper and magazine production, especially for advertisements, is now most common.

One special application of silver photocopying is the automatic high-speed Listomatic Camera, developed in 1953. This machine photographs data on tabulating cards and prints it in columnar form on roll film; the film negative is then used for printing by photolithography (Fig. 2). The process is particularly suited to the preparation of directories and other lists.

In sum, the silver halide technique can produce copies having excellent range of contrast, resolution, superb continuous tone, and the permanence of rag paper. Relatively speaking, it is not inexpensive, but its quality has not yet been matched by any of the five processes below.

**Transfer processes.** Transfer processes are of the soft gelatin (or Verifax) type and the diffusion type. Both of these reflex contact processes produce copies that are generally less expensive than copies by the silver process, but more expensive than those from any of the other processes that follow; and their quality is better than thermographic, diazo, or electrostatic prints. However, they are not good for tone, and the image will deteriorate over a period of years. Both processes are packaged in a variety of compact machines utilizing a monobath to combine developing and fixing while retaining, to a marked degree, the sensitometric characteristics of conventionally developed paper.

The gelatin type, introduced in 1952, is a physical transfer of dye from a gelatin emulsion suspending silver halide, a dye-forming component, and a hardening agent. The Verifax matrix is processed in a monobath solution in which the black dye is formed throughout the gelatin, but is hardened only

in the background areas. The wet matrix is pressed against plain paper to get the positive copy from the unhardened dye. By pressing additional sheets, from two to nine additional positives may be obtained, each fainter than the preceding; the process can also produce masters for offset duplicators and translucent copies for diazo prints.

The diffusion type, introduced in 1950, uses a silver chloride emulsion on film or paper, with the image diffused to opaque paper, translucent paper, or clear film. The light-sensitive negative paper, after exposure with the original, is matched with a chemically coated positive paper which is not light-sensitive. They pass through a monobath, are squeezed together for 15–30 sec, and are then peeled apart. The unexposed text area of the negative gives up its unused silver salts to the positive paper, thereby coming in contact with the chemical coating of the positive paper to form a black image. Generally, more than one positive per negative is not possible. Although more than 60 different models using this process are marketed, the essential differences are in the method of holding the original material to be copied, tightly bound books being the major problem because it is difficult to make firm contact with the light source at the inner margin. The process can also produce a two-sided positive copy on airmail tissue, on translucent stock, or on transparent film. In 1959, a projection camera was made available with a higher-speed sensitized negative, providing this process with the ability to change the size of the original.

**Thermography.** This is not a photographic process, because instead of light, it uses the heat of infrared rays for exposure. Yet this process, available since 1950, is a contact reflex rapid-copy printing process, similar in its use to the transfer process. It is a direct-positive process—no negative is created. The original copy must have carbon or a

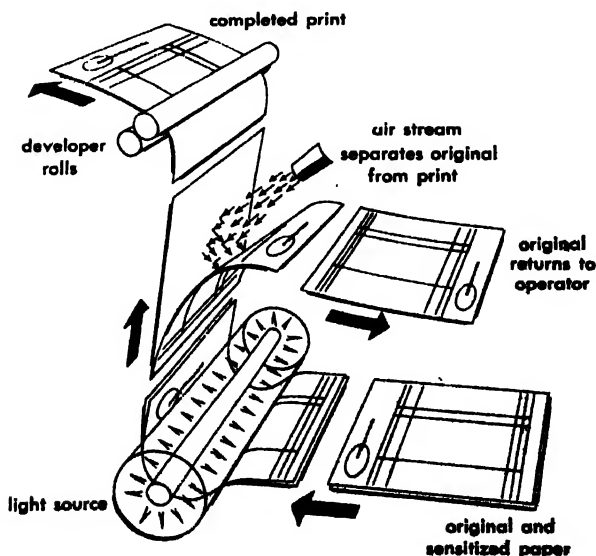


Fig. 3. Flow diagram of Whiteprint machine. (Charles Bruning Co.)

metallic compound in the text ink to transform the radiant energy to heat and so effect the desired chemical change on the substance laminated between the transparent sheet of paper and the white waxy backing. The heat-sensitive substance undergoes chemical change and produces a black image lacking maximum sharpness. It is a completely dry operation completed in about 4 sec; finished copies remain sensitive to heat and can become increasingly dark. It does not satisfactorily handle tone, and the print is not as sharp as copies by most other processes because of the difficulty of focusing the long wavelengths of heat. For inexpensive, short-use, clean, and rapid copying of correspondence or printed textual matter for informational purposes, this process excels.

**Plan-copying.** This is a simple contact operation using a number of possible chemical processes to print from a translucent original, or by direct-positive paper from opaque originals. Materials used are so insensitive to light that powerful arc lamps must be used for exposure. There is less of the usual wastage of materials in this process than in others. The photocopying machine may be a glass tube, box, or rotary-drum device with a large surface to accommodate architectural plans, engineering drawings, charts, maps, or other such material. Of the many processes, the three following are common.

**Brownprint** (also called *sepia negative* or *vandyke*) is an intermediate for making prints, introduced in 1895. The material is paper sensitized with ferric iron and silver salts, the first ingredient being the light-sensitive material. Exposure reduces the iron salt, and when developed by immersion in water, the ferrous salt reduces the silver salt to metallic silver. Washing with a hypo solution removes any unreduced silver and leaves white lines on a brown background. From this negative, a brown line print can be made, or a Phototracing can be made on paper having a wash-off silver gelatin emulsion the image of which can be erased with a wet eraser and additions made in ink.

**Blueprint** (also called *cyanotype*), dating from 1842, is a ferropussiate paper, sensitized with a mixture of ferric salt and potassium ferricyanide, developed by immersion in water. The result is white lines on a background of Prussian blue. As with the brownprint, the color of line and background can be reversed by printing from a translucent negative so as to make a blue line print.

**Whiteprint** (also called *dyeline* or *diaz print*), dating from the 1920s, is produced on diazo paper, or film, the emulsion containing a diazo compound and a coupling or activating component. The process is based on sensitivity to ultraviolet light, and development is by ammonia vapors or a liquid application (Fig. 3). The use of this method to produce translucent film originals from which multiple prints can be inexpensively made on paper has been a highly developed technique. Indeed, despite the fact that the image will deteriorate somewhat over a period of years, this process has largely super-

seded the familiar blueprint chiefly because of better appearance, easier use for notations, somewhat better quality and comparable cost.

It should be added that small diazo-process machines are available to copy letter-size materials and pages from bound books. When limitations imposed by the material to be copied can be overcome, the excellent sharpness of a diazo print and its exceedingly low cost make it competitive with the transfer process.

**Electrostatic processes.** There are three distinct dry photoelectrical processes producing positive copies without a negative intermediary: xerography was invented in 1937; Electrofax was announced in 1954; and Smokeprinting was patented in 1958.

Xerography is a printing method using a photoconductive plate having an electrically conductive backing material coated with vitreous selenium. When the plate is precharged to a 6000-volt screen potential by a corona discharge which imparts a uniform electrostatic charge, the coating becomes sensitive to light and the charge is dissipated to a ground by light rays reflected from the white parts of the document being copied. The sensitized plate may be exposed by contact, by projection, or in a camera to achieve enlargement or reduction. The latent image is the remaining positive charge which attracts negatively charged black powder (a mixture of a carrier and a resinous pigment) which is then heated and passed to paper, on which it is fused. A resolution adequate for most textual matter can be achieved, and recent developments have improved the handling of continuous tone. Experimental work has been done on the use of color and in the depositing of images on copper laminate and on clear acetate sheets for lantern slides. Besides its use for enlarging microfilm, some of its important uses are for making translucent negatives for diazo printing, paper and metal plates for offset lithography, and masters for spirit duplicating (Fig. 4).

Electrofax is similar to the xerographic process except that it substitutes an electrically charged paper on which the copy is printed for the selenium-coated plate. The paper is coated with a thin layer of special zinc oxide in a resin binder, and it is sensitive to light only after having been given a negative electrostatic charge upon entering the machine; it can therefore be stored without deterioration for long periods before being used. The finished print is exceedingly stable. Experimental work has been done on many applications similar to those of xerography, as well as on satisfactory treatment of half-tone and continuous tone, production of relief printing plates, use as a dry offset process, and on electronic typesetting at 2000 characters per second under control of magnetic tape or punched paper tape. The process appears to be intrinsically more flexible than does xerography.

Smokeprinting is a process which deposits electrically charged particles on paper or other material. The paper is held behind a sheet of glass

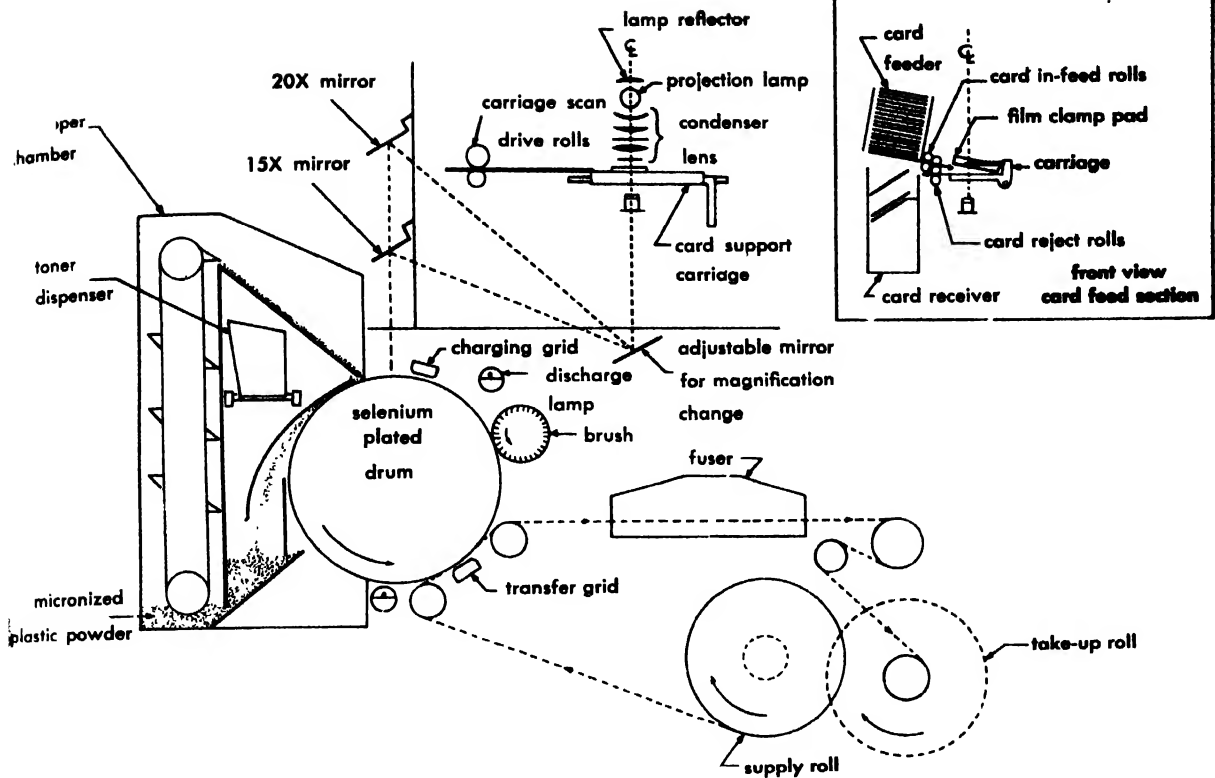


Fig. 4. Schematic diagram of continuous printer of microfilm enlargements. (Inset) Feed mechanism for

unitized microfilm mounted on aperture cards. (Haloid Xerox)

which is backed with a thin metallic coating. The mist of particles is dispensed from behind the paper by an electrode which gives it a charge. A positive or negative print can be made, depending upon the smoke material used and whether it is charged positively or negatively. Experimental work has been successful on a Vertical Step-and-Repeat Microfilming Camera for producing microfilm images on sheet film and on the Photronic Reproducer for microfilm enlargement.

**Electrolysis.** Electrolysis has been applied to photocopying under a process developed in 1957. The exposure is made by projection on a sheet consisting of a paper support, a thin aluminum laminate, and a coating of a white photoconductive substance. A dc potential applied across the electrolyte and the aluminum underlayer are necessary for development, which is through contact with the electrolyte and thereby produces the image on the exposed surface by electrolysis. A positive image is produced from a negative microfilm projection because the light exposure causes the photoconductor to lose resistance, resulting in electrolysis taking place in exposed rather than in nonexposed areas.

**Microfilming.** This is a documentary reproduction process. The first patent for a commercial application was granted in 1859, although wide application dates from 1928. A microfilm copy may be defined as a transparent photocopy at a reduction sufficient that optical enlargement is required for normal reading and with resolving power sufficient

for accurate recording of textual and tonal detail. Normal reductions are from 12 to 22 diameters, although 100 diameters is feasible. Resolving power is well over 100 lines/mm—generally about 145 lines/mm for good positive film, 160 lines/mm for good microfilm lenses, and 180 lines/mm for good negative film. Emulsions giving 500 lines/mm are being developed. Such a microphotograph is distinguished from a macrophotograph, which is a copy near the original size as in all the processes described above.

The microfilm is composed of a slow-speed panchromatic emulsion on a mechanical base of cellulose acetate (nitrocellulose has unacceptable qualities). In the United States, it generally is no longer used with perforated edges in order that better use may be made of the film width. Positive copies may be made from silver negatives by contact printing on silver or Kalfax film, or sometimes on paper. Diazo film is used for exact duplication of silver roll film, a negative-to-negative or positive-to-positive print. These three emulsions are available in the common 35-mm roll form, and also in rolls of 16 mm for material such as bank checks which have no fine detail, 70 and 105 mm for material such as engineering drawings, and in flat sheets of 7½ by 12½ cm and larger sizes which have filing and searching advantages.

Kalfax microfilm is a variant use of diazo which on exposure creates gas bubbles which form the image by scattering the light. Kalfax (announced in



1955) is a plastic emulsion on a polyester base. Upon exposure to ultraviolet light, the photosensitive compound in the background area decomposes in the thermoplastic vehicle, with one of the products being nitrogen gas. When the film is developed by heat (for 2 sec at 255°F), the high pressure created by the gas blows microscopic air sacs; and fixation by ultraviolet light stabilizes the compound by permitting the nitrogen now created in the text area to diffuse out of the emulsion in about 8 hours. The air sacs serve to scatter the light falling on them during projection for reading. Where they do not exist to scatter the light, the compound casts a shadow which forms the image.

Microfilm is very inexpensive of materials, is valued as a substitute for deteriorating paper and for the space and weight it saves over full-size copies, is advantageous as a flexible intermediate, and has the physical properties of other silver and diazo copies. In any application, such as those described below, detail will be lost increasingly the more distant the generation is from the original (that is, original copied to negative copied to positive copied to a second negative equals three generations removed, this loss being estimated as roughly 30% in each generation).

Equipment for microfilming is specialized because of the exacting requirements. Cameras have exceptionally fine lenses, and may have a flat bed or rotary feed and be manual or automatic in operation. Similarly, processing equipment varies from hand-fed deep tanks to large automatic machines adapted from those designed for motion picture film. Reading machines are also available in considerable variety.

Besides the advantages of microfilm in its own right, other applications and specialized techniques have brought microfilm to a high state of technical development. These may be classed as enlargement techniques, publishing techniques, and data-processing applications.

Enlargement printing, both individually of single frames and automatically and continuously of rolls of frames, can be accomplished by any projection process. Enlargement is possible from frames selected while viewing microfilm on a reading machine by the inclusion in the machine of a device for producing an electrolytic copy or a silver halide copy developed in a monobath. The most advanced automatic enlarging machine is the Copyflo, which was first available in 1956. This xerographic enlarger makes photocopies from negative or positive 16-mm or 35-mm microfilm, as well as from original documents, at the rate of 20 ft/min on rolls of unsensitized paper up to 11 in. wide.

Publishing of opaque microtexts is accomplished by three variant edition processes. Microcards, proposed in 1944, are photographic prints  $7\frac{1}{2} \times 12\frac{1}{2}$  cm in size, prepared from 16-mm or 35-mm film, commonly at a reduction of 20 diameters, with indexing data legible to the naked eye at the top of the card. Readex Microprint is a somewhat similar product on  $6 \times 9$ -in. cards, but is prepared from

microfilm and printed by offset. Microlex is similar to the Microprint in appearance, but is a photographic print from a sheet negative (microfiche) made on a step-and-repeat camera, and two positives containing consecutive pages are laminated back to back. Microfiche is the name applied to transparent forms of microtext in various sizes of sheet film; it is not an edition process.

Data processing has used microfilm in the unitized, strip, and roll form. Unitized film is roll film cut and handled in units of single frames, whereas a film strip is roll film handled in lengths of 2-10 frames; diazo film is commonly used for making copies. Both unitized and strip film are commonly mounted in or on cards having indexing information which is readable to the naked eye, and if the card has a rectangular hole through which the film may be viewed, it is called an aperture card. Two versions using a different technique are Microstrip, which is an opaque paper strip printed from microfilm and having a moisture-type adhesive on the back, and Microtape, which is similarly an opaque strip but has a pressure-sensitive adhesive. Much special equipment is available for mounting, viewing, enlarging, and duplicating film mounted on aperture cards; and use of such cards with tabulating machines has become a common application of microfilm. In unitized form, film has reached its most sophisticated application in the Minicard system, which automatically searches units of film, 35 mm by 16 mm in size, containing both textual matter reduced at 60 diameters and digital information for photoelectric eye selection. Roll film is also being applied in data processing systems; an early device was the Rapid Selector developed from principles suggested by Vannevar Bush in 1945, and a more highly developed machine is the FLIP (Film Library Instantaneous Presentation) which, upon instruction from a keyboard, punched cards, or magnetic tape, automatically locates and projects for viewing coded frames within 1600 feet of film. See PRINTING. [D.C.W.]

**Bibliography:** H. W. Ballou, *Guide to Microreproduction Equipment*, 1959; C. M. Lewis and W. H. Offenhauser, *Microrecording; Industrial and Library Applications*, 1956; *Manual on Document Reproduction and Selection*, FID Publ. 264, 1953 and continuation; H. R. Verry, *Document Copying and Reproduction Processes*, 1958.

## Photodiode

A semiconductor two-terminal component with electrical characteristics that are light-sensitive. All semiconductor diodes are light-sensitive to some degree, unless enclosed in opaque packages, but only those designed specifically to enhance the light sensitivity are called photodiodes.

Most photodiodes consist of semiconductor  $p$ - $n$  junctions housed in a container designed to collect and focus the ambient light close to the junction. They are normally biased in the reverse, or blocking, direction; the current therefore is quite small in the dark. When they are illuminated, the current

is proportional to the amount of light falling on the photodiode. For a discussion of the properties of  $p$ - $n$  junctions, see JUNCTION DIODE.

Photodiodes are used both to detect the presence of light, and to measure light intensity. See PHOTO-ELECTRIC DEVICES. [W.R.SI.]

## Photoelasticity

An experimental technique for the measurement of stresses and strains in material objects by means of the phenomenon of mechanical birefringence. Photoelasticity is especially useful for the study of objects with irregular boundaries and stress concentrations, such as pieces of machinery with notches or curves, structural components with slits or holes, and materials with cracks. The method provides a visual means of observing over-all stress characteristics of an object by means of light patterns projected on a screen or photographic film. Regions of stress concentrations can be determined in general by simple observation. However, precise analysis of tension, compression, and shear stresses and strains at any point in an object requires more involved techniques. Photoelasticity is generally used to study objects stressed in two planar directions (biaxial), but with refinements it can be used for objects stressed in three spatial directions (triaxial).

For biaxial studies, a model geometrically similar to the object to be analyzed is prepared from a sheet of special transparent material and loaded as the object would be loaded.

**Use of birefringent phenomenon.** Model materials commonly used for photoelasticity are Bakelite, celluloid, gelatin, synthetic resins, glass, and other commercial products that are optically sensitive to stress and strain. The materials must have the optical properties of polarizing light when under stress (optical sensitivity) and of transmitting it on the principal stress planes with velocities dependent on the stresses (birefringence or double refraction). In addition, the material should be clear, elastic, homogeneous, optically isotropic when under no stress or strain, and reasonably free from creep, aging, and edge disturbances.

When the stressed model is subject to monochromatic polarized light in a polariscope the birefringence of the model causes the light to emerge refracted into two orthogonal planes (see POLARIZED LIGHT). Because the velocities of light propagation are different in each direction, there occurs a phase shifting of the light waves. When the waves are recombined with the polariscope, regions of stress where the wave phases cancel appear black, and regions of stress where the wave phases combine appear light. Therefore, in models of complex stress distribution, light and dark fringe patterns (isochromatic fringes) are projected from the model. These fringes are related to the stresses.

When white light is used in place of monochromatic light, the relative retardation of the model causes the fringes to appear in colors of the spec-

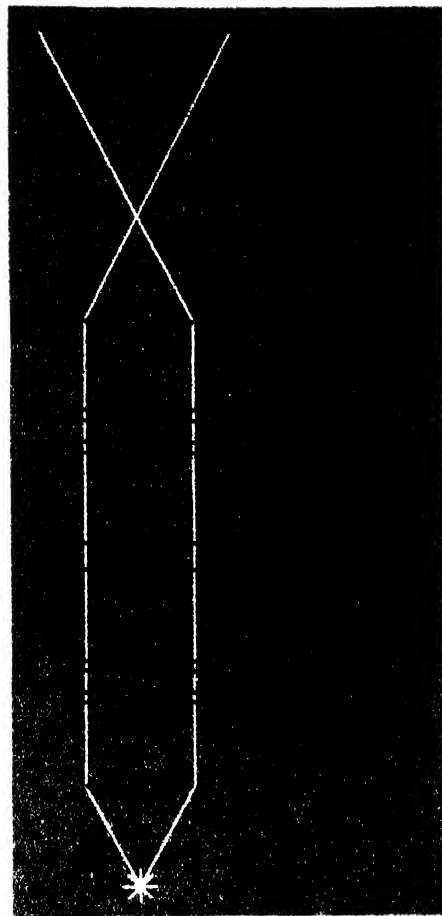


Fig. 1. Basic photoelastic polariscope.

trum. White light is often used for demonstration, and monochromatic light is used for precise measurements.

**Polariscope.** A basic polariscope used in photoelasticity has a light source (generally monochromatic), a collimating lens, a polarizer, and a quarter wave plate (Fig. 1). This plate is a birefringent material that causes the relative retardation of light to be exactly one-quarter the wavelength of the light. Next in the optical path is a planar model of the object under test and stressed in the direction of the plane. Finally there are a second quarter wave plate, a polarizer called the analyzer, a focusing lens, and a viewing screen or film. Many variations of this basic transmission-type apparatus are in use. Other lenses may be added and components rearranged. If appropriate mirrors are added, the polariscope converts to a doubling type which is useful for the study of thin models under low stress, as the number of fringes doubles.

A typical isochromatic fringe pattern shows the effect on a flat plate with a central hole, pulled at the upper and lower ends (Fig. 2). The congregation of fringes at the boundary of the hole indicates a region of stress concentration, typical of stress behavior at cutouts. To study the exact stress at a given point in the model, the model is gradually loaded (from a condition of no load) and the num-

ber of fringe changes (fringe order) at that point is observed. Special equipment is sometimes employed to obtain partial fringe orders and to sharpen vague fringe boundaries. The fringe order is directly related by a calibrated constant to the difference of the principal stresses at that point.

**Determination of principal stresses.** Shear stresses can be related mathematically to the difference of the principal stresses, thereby relating shear stresses directly to fringe order. High shear stresses often cause the material to yield or fail, so that a point of large fringe order indicates a point of potential failure. In many applications of photoelasticity a knowledge of shear stress is all that is needed. This fact makes photoelasticity a simple and direct tool for investigation of complex stress systems. However, if principal stresses and their directions are required, additional experimentation is necessary, as described later.

Isoclinic fringes are a different set of interference patterns made by using white light, removing the quarter wave plates and rotating the polarizer and analyzer a fixed number of degrees. These fringes represent lines making known angles with the principal planes of stress.

Stress trajectories are lines of principal stress directions over the model, obtained graphically from the isoclinic fringes. Stress trajectories are not lines of constant stress.

The determination of the principal stresses requires additional information, which may be obtained in several ways. Principal stresses are determined analytically by differences of the shear stresses based on equilibrium equations. This procedure requires a numerical point by point study of the model, utilizing the shear stresses and stress trajectories. Principal stresses can be found analytically or experimentally by solution of Laplace's equation of elasticity (see ELASTICITY). In principle, this procedure supplies equations pertaining to the sum of the principal stresses at any point in the model. Utilizing equations for the difference of the principal stresses from the isochromatic fringe orders, the stresses may be found by solving the two equations for the two principal stresses. Principal stresses are found experimentally by measuring the changes in thickness of the model under stress. Because thickness changes caused by the Poisson's ratio effect are minute, a sensitive measuring device such as an optical interferometer is needed, although direct-reading thickness gages are sometimes used. The interferometer produces fringe patterns called isopachic fringes. This method essentially provides information regarding the sum of the principal stresses as with Laplace's equation. Another experimental method is to pass polarized light obliquely to the surface of the model. The relative retardations of the light produce interference fringes. These oblique fringe orders can be related to the principal stresses differently from those obtained by isochromatic fringes. Using the information on stresses from the isochro-

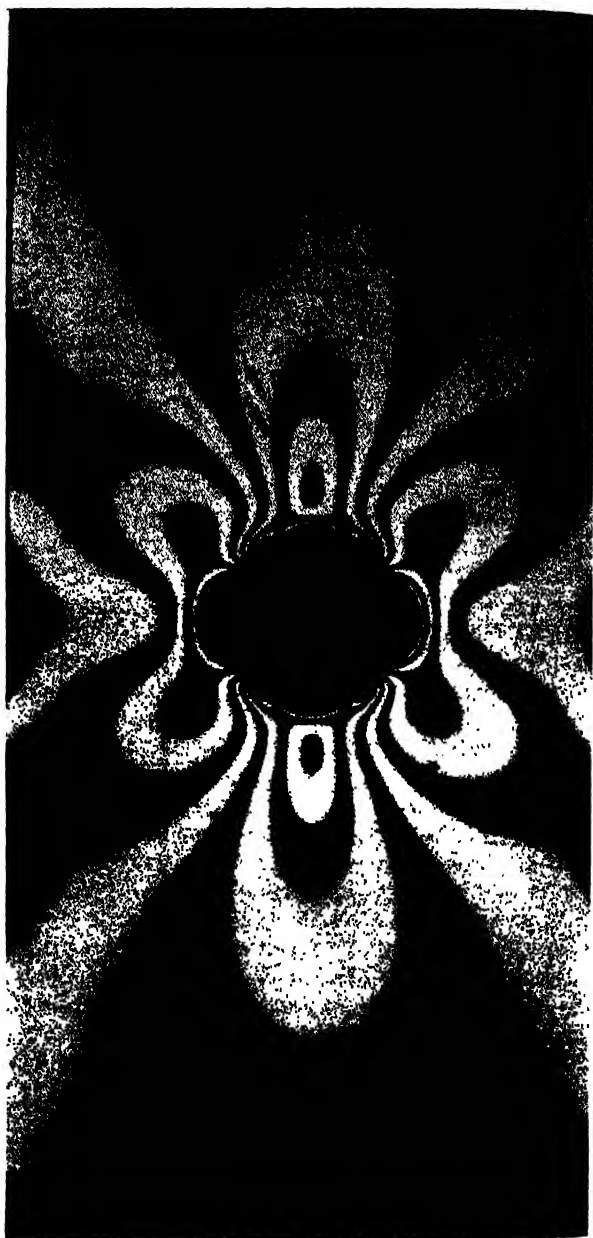


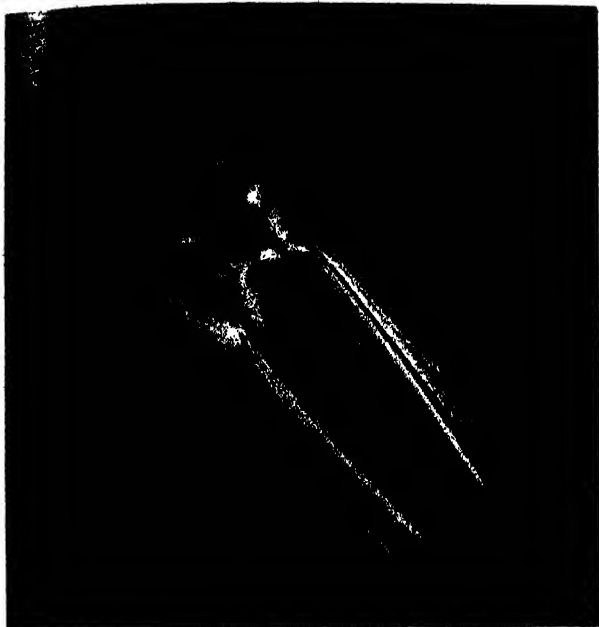
Fig. 2. Isochromatic fringe pattern for plate with hole. (From M. M. Frocht, *Photoelasticity*, vol. 2, Wiley, 1948)

matics in conjunction with the oblique relations, the principal stresses may be obtained.

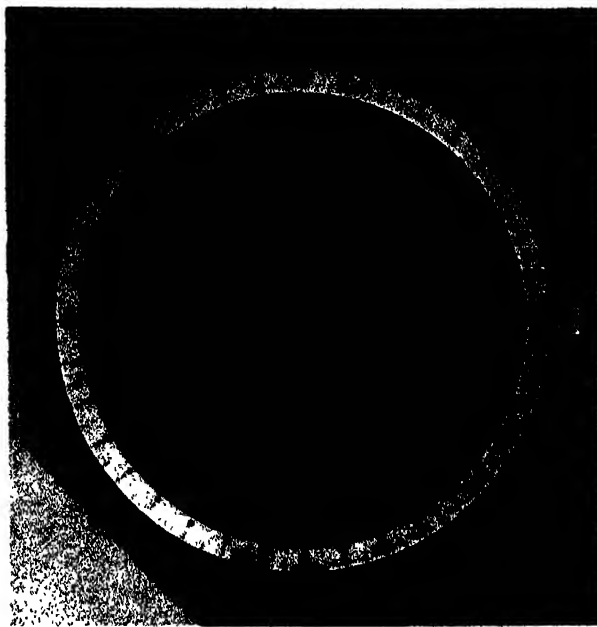
With care, stresses determined by photoelasticity are 98% accurate. With stresses determined, strains may be easily computed by elastic relations.

**Three-dimensional measurements.** Three-dimensional photoelasticity is also possible, although the techniques and stress-strain relationships are more involved than for planar objects.

The frozen stress method is well suited for three-dimensional studies. Certain optically sensitive materials, such as Bakelite, when annealed in a stressed condition retain the deformation and birefringent characteristics of the initially stressed state when the load is removed. A three-dimen-

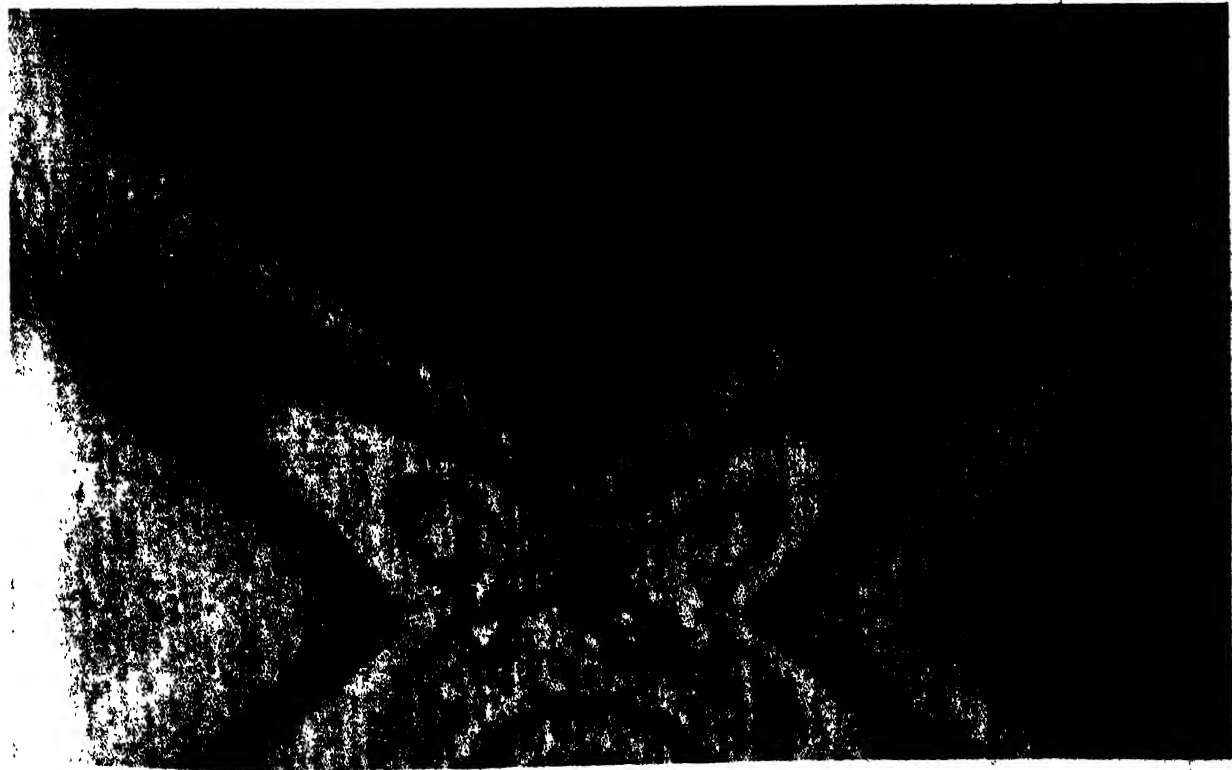


Typical stress patterns in a plastic model viewed under polarized light. (Bausch & Lomb Optical Company)



Stress concentrations in plastic models of roots of steam turbine blades under simulated operating conditions. (Westinghouse Electric Corporation)

Stress distribution around a circular hole and two saw cuts using photostress technique. Plastic photoelectric coating on actual steel specimen allows study of actual metal part rather than of a plastic model. The proximity of color fringes in the vicinity of the saw cuts indicates high stresses in that region. (Battelle Memorial Institute)





sional model may therefore be cut up into slices. These slices may then be analyzed individually on somewhat the same principles as are the planar models.

The scattered light technique may also be used for three-dimension models, such as torsion bars. The scattered light principle is based on the fact that polarized light passing through birefringent materials scatters in a predictable manner, acting as an optical analyzer in a polariscope.

**Measurements on actual objects.** The PhotoStress method is essentially a variation of the normal polariscope. A sheet of birefringent material is bonded to the polished surface of the actual object to be studied, with a polarizer and quarter wave plate interposed between the light and the object. The light (usually white light) passes through the stressed birefringent material and is reflected back through the quarter wave plate and the polarizer (which now acts as an analyzer). Isochromatic fringes are projected as in a normal polariscope. Dependent on good bonding between the birefringent material and the stressed object, PhotoStress has the advantage that it can be used on actual structures under service loads without being reduced to model form. It may be used to find surface stresses and strains on curved or three-dimensional objects. Although photoelasticity is generally limited to elastic behavior, PhotoStress may be used to determine the nonelastic strains (but not stresses) of objects, provided the bonded optical material remains elastic. The technique may also be used with nonheterogeneous materials such as wood and concrete.

Dynamically induced stresses and strains may also be studied by photoelasticity when high-speed motion picture cameras are used to photograph the fringe patterns. See STRESS AND STRAIN. [W.Z.]

**Bibliography:** E. G. Coker and L. N. G. Filon, *A Treatise on Photo-Elasticity*, 2d ed., 1957; M. M. Frocht, *Photoelasticity*, 2 vols., 1941-1948; M. I. Hetényi (ed.), *Handbook of Experimental Stress Analysis*, 1950; G. H. Lee, *An Introduction to Experimental Stress Analysis*, 1950.

## Photoelectric devices

Devices in which a significant and useful change in electrical characteristics is caused by incident radiation energy. While all materials are sensitive to or react with electromagnetic radiation to some degree, photoelectric devices exhibit a significant sensitivity which can be used to detect, measure, or convert the incident radiation. Photosensitivity herein implies the infrared and ultraviolet sections of the spectrum as well as the visible. See PHOTOELECTRICITY.

Photoelectric devices react with the incident radiation in a variety of ways. In photoconductive or photoresistive cells the resistance varies depending upon the intensity and wavelength of the radiation; these cells are particularly sensitive in the infrared region of the spectrum (see PHOTO-

CONDUCTIVE CELL). Photovoltaic cells, sometimes called photonic or boundary layer photocells, generate an output voltage when excited by photons of light. A variety of photovoltaic cells are available to cover sections of the infrared and ultraviolet regions, as well as the visible. When loaded with a low impedance of about 1000 ohms or less, the ratio between the light input and current output is linear and the frequency response of the system is fairly flat to about 2 kilocycles per second. When the cell is open circuited, the voltage output increases logarithmically with light input. See PHOTOVOLTAIC CELL.

Photodiodes and phototransistors are photovoltaic devices in which the junction barrier height is modulated by incident light. In the photodiode this causes the reverse current to vary as a function of the light intensity and back voltage, while in a transistor it can be employed as the input or injection signal to a fixed-biased circuit or used to vary the bias conditions. See PHOTODIODE; PHOTOTRANSISTOR. [D.B.K.]

## Photoelectricity

Visible light or other electromagnetic radiation incident on a solid, liquid, or gas can liberate electric charge which moves in an electric field; this process is called photoelectricity. The term includes three distinct phenomena, as follows:

1. In the external photoelectric effect, often called photoemission and first explained on a quantum basis by Albert Einstein, electrons are ejected from a solid (or liquid) surface into a surrounding vacuum; the common multiplier phototube depends upon this effect.

2. In a gas, electrons and positive ions may be produced in the process known as photoionization. An application of this is made in the use of ionization chambers to detect x-rays.

3. Mobile electrons and positive "holes" which remain inside a solid may give rise either to photoconduction or to the photovoltaic effect. Photoconduction is used in television camera tubes, in electrical duplicating processes, and in control devices (where a simple external battery furnishes the electric power). The photovoltaic cell is an energy-conversion device which furnishes its own power. It is used, for example, in photographic exposure meters and in solar batteries.

In addition to the three phenomena just described, there is an internal photoelectric process in an atom called the Auger effect, or autoionization. A part of the absorption of x-rays can be attributed to the Auger effect, which involves photoelectric absorption of the x-ray quanta with resulting ejection of electrons from the atom. X-ray absorption also occurs by photoelectric excitation and by fluorescence.

Finally, there is the inverse photoelectric effect, which is simply the inverse of the normal photoelectric process. In this effect, an electron is absorbed by a solid and a photon emerges. The inverse pho-

toelectric effect is not commonly investigated because it is an extremely difficult process to measure. See AUGER EFFECT; PHOTOCONDUCTIVITY; PHOTOEMISSION; PHOTOVOLTAIC EFFECT; *see also* COMPTON EFFECT. [L.A.]

## Photoemission

Photoemission, also called the external photoelectric effect, is the ejection of electrons from a solid (or less commonly, a liquid) by incident electromagnetic radiation. The visible and ultraviolet regions of the electromagnetic spectrum are most often involved, although the infrared and x-ray regions are also of interest. For important practical applications of photoemission, *see* PHOTOTUBE; TELEVISION CAMERA TUBE.

The salient experimental features of photoemission are the following: (1) there is no detectable time lag between irradiation of an emitter and the ejection of photoelectrons; (2) at a given frequency, the number of photoelectrons ejected per second is proportional to the intensity of the incident radiation; and (3) the photoelectrons have kinetic energies ranging from zero up to a well-defined maximum, which is proportional to the frequency of the incident radiation and independent of the intensity.

**Einstein photoelectric law.** These characteristics can not be explained by J. C. Maxwell's theory of electromagnetic waves. In 1905 Albert Einstein made the clarifying assumption that the radiation had characteristics like those of particles when it delivered energy to electrons in the emitter. In Einstein's approach, the light beam behaves like a stream of photons, each of energy  $h\nu$ , where  $h$  is Planck's constant, and  $\nu$  is the frequency of the photon (Fig. 1). The energy required to eject an electron from the emitter has a well-defined minimum value  $\phi$  called the photoelectric threshold energy. When a photon interacts with an electron, the latter absorbs the entire photon energy. *See* PHOTON.

For  $h\nu$  values below the threshold, photoelectrons are not ejected. Even though the electrons absorb photon energy, they do not receive enough to surmount the potential barrier at the surface, which normally holds the electrons in the solid. (For a discussion of the surface potential barrier, *see* SCHOTTKY EFFECT.) The threshold energy  $\phi$  is associated with a threshold frequency  $\phi/h$  and

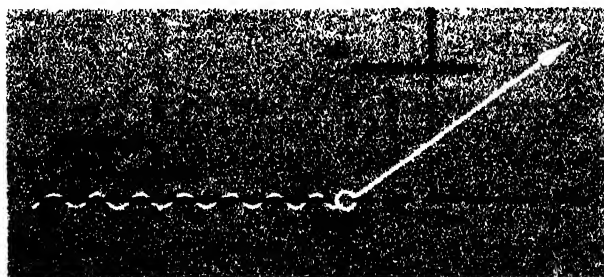


Fig. 1. External photoelectric effect.

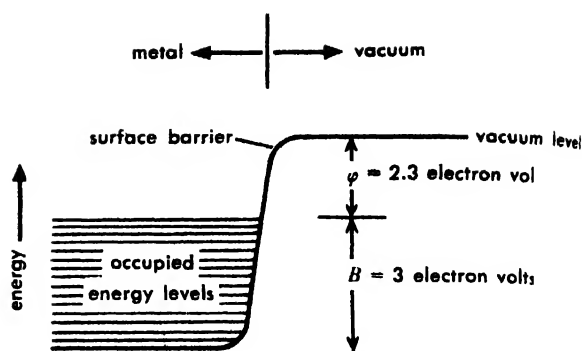


Fig. 2. Energy diagram for electrons in sodium. The photoelectric threshold energy is  $\phi$ ; in a metal  $\phi$  is equal to the electronic work function. The band of energy levels occupied by almost free electrons has a width  $B$ .

a threshold wavelength  $ch/\phi$ , where  $c$  is the velocity of light. For photon energies above  $\phi$ , the kinetic energies of photoelectrons range from zero up to a maximum value,  $E = h\nu - \phi$ . This is the Einstein photoelectric law, and  $E$  is commonly termed the Einstein maximum energy. Careful photoelectric experiments by R. A. Millikan in 1916 fixed  $h$  in Einstein's law with considerable precision and furthered its identification with the constant which M. Planck had used in his theory of black-body radiation.

**Metals.** The Einstein law is based only on the photon hypothesis and on the conservation of energy. It does not take into account momentum, which must also be conserved. The incident photon has a momentum  $h\nu/c$  which is negligible compared to the change in momentum of the electron when it gains the energy  $h\nu$ . Thus, it is not possible for a free electron to absorb the entire energy of a photon. In order for this to happen, the electron must be bound to another body, which takes up the recoil momentum. *See* COMPTON EFFECT.

Figure 2 shows an energy diagram of the electrons in the metal sodium. There is a potential barrier at the surface, which the electrons must surmount before they can escape. The most easily ejected electrons must acquire 2.3 eV of additional energy from photons in order to do this. This 2.3 eV is the electronic work function, which for a metal is equal to the photoelectric threshold energy. *See* WORK FUNCTION (ELECTRONIC). Inside the metal the electrons occupy a band of energies about 3 eV wide. These electrons are said to be quasi-free. This means that they behave in many ways like a gas of free, noninteracting electrons; nevertheless they move in the periodic potential due to the positive sodium ions, and in this sense, they are bound. *See* FREE-ELECTRON THEORY OF METALS.

In this situation, two types of photoemission are theoretically possible, the surface effect and the volume effect. In the surface effect, recoil momentum is communicated to the crystal because the electron is coupled to the barrier at the surface during photon absorption. In the volume effect, the



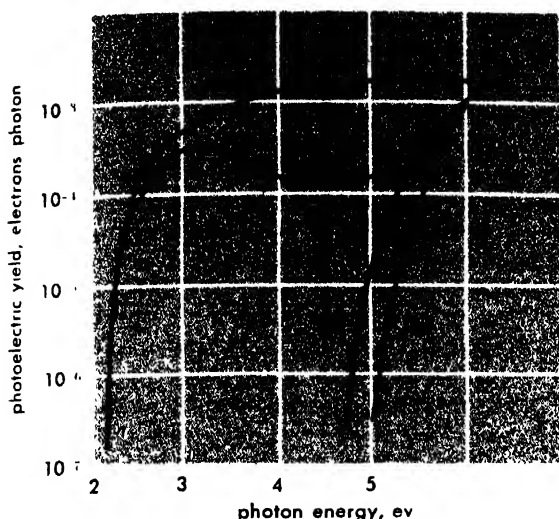


Fig. 3. Spectral distribution of the photoelectric yield from typical samples of barium (Ba), potassium (K), platinum (Pt), and tellurium (Te). Platinum and tellurium have practically the same electronic work function. Note the higher threshold and more steeply rising curve for tellurium, which is a typical elemental semiconductor.

electron is coupled to the internal periodic potential.

Experimental determination of the relative importance of surface and volume photoeffects in metals is difficult. Experiments by H. Mayer and his collaborators indicate that for potassium the volume effect is predominant for photon energies from the threshold value at 2.1 eV to at least 4 eV. Thus, the photoelectric emission increases as the thickness of a potassium film increases. (This would not be true for the surface effect.) Photoelectrons can escape from depths greater than  $10^{-6}$  cm when excited by light in the threshold region.

Thus far the photoelectric threshold has been treated as a sharply defined quantity. This is precisely true for metals only at temperatures near absolute zero. At higher temperatures, the upper edge of the band of occupied electron energy states in Fig. 2 is no longer sharp. It becomes diffuse because of thermal agitation. Electrons may be then emitted for photon energies less than the threshold value,  $\phi$ . At ordinary room temperatures, for example, measurable photoemission appears for photon energies as much as 0.2 eV below the threshold. R. Fowler has developed a convenient graphical technique, known as a Fowler plot, for determining the absolute-zero threshold from data taken at higher temperatures on the spectral dependence of the photoelectric yield, which is the number of photoelectrons ejected per incident photon. L. A. DuBridge has developed a similar technique using either the temperature dependence of the photoelectric yield or the distribution of photoelectrons in energy at fixed frequency. These treatments show that the photoelectric yield is approximately

proportional to the quantity  $(h\nu - \phi)^2$  when the photon energy  $h\nu$  is within about 1 eV of the threshold energy  $\phi$ . Figure 3 shows a graph of the spectral dependence of photoelectric yield for some typical emitters. Figure 4 shows typical energy distributions. Photoelectric yields from metals are of the order of  $10^{-3}$  electron per incident photon when  $h\nu - \phi$  is 1 eV. Photoelectric threshold energies range from 2 eV for cesium to values such as 5 eV for platinum. They vary for different types of crystal faces on the same crystal and are exceedingly sensitive to small traces of adsorbed gases.

**Semiconductors.** The photoelectric behavior of semiconductors such as germanium or tellurium differs from that of metals. As shown in Fig. 5, the electrons in a semiconducting emitter completely occupy a closed band of energies, which lies just below a so-called forbidden energy band (see BAND THEORY OF SOLIDS). The electrons behave quite differently from those in metals. As a result, the photoelectric threshold energy  $\phi'$  is larger than the electronic work function  $W$ . Thus, a semiconductor exhibits a higher photoelectric threshold energy than a metal having the same work function. An example of this is shown for the metal platinum and the semiconductor tellurium in Fig. 3. Both this particular platinum sample (Pt) and the tellurium (Te) have the same electronic work function, about 4.8 eV. The photoelectric threshold of the platinum is equal to the work function, whereas that for the tellurium is clearly higher. Spectral and energy distributions are shown in Figs. 3 and 4. Clean surfaces of silicon, germanium, and cer-

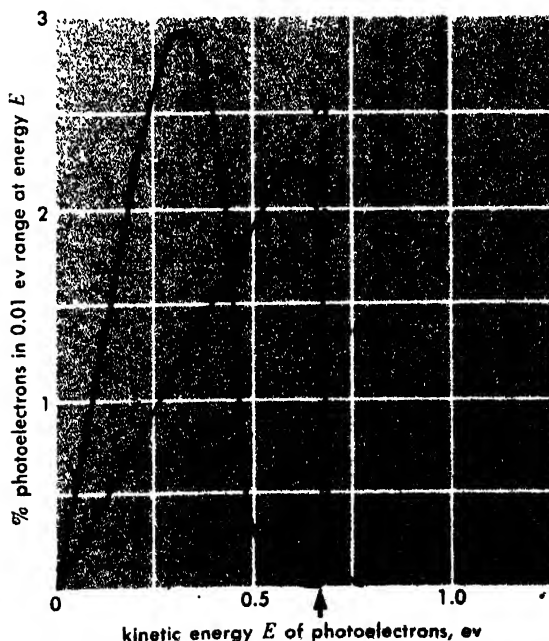


Fig. 4. Energy distributions of photoelectrons from tellurium and from a metal having the same work function. The solid line for the metal shows results for ordinary room temperature, and the dashed line is for absolute zero. The arrow marks the Einstein maximum energy. The photon energy is 5.42 eV.

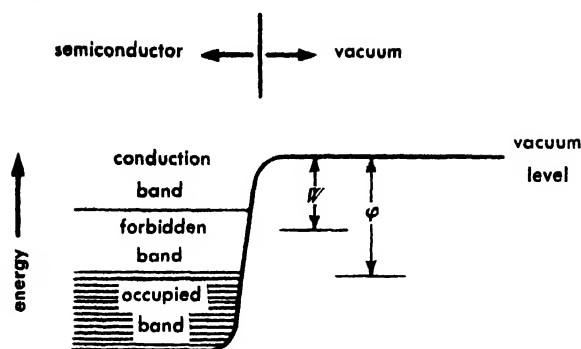


Fig. 5. Energy diagram for electrons in a semiconductor. The occupied band is filled with bound electrons that behave differently from those in metals. As a result, the electronic work function  $W$  is smaller than the photoelectric threshold energy  $\phi$ .

tain semiconducting chemical compounds have been made by cleaving single crystals in ultra-high vacuum. Both the surface photoelectric effect and the volume effect have been measured. From the measurements of the volume effect, valuable information on the detailed nature of the electron energy bands has been deduced from structure that occurs in photoelectron energy distributions.

A particularly interesting and important kind of photoemitter is typified by cesium antimonide,  $\text{Cs}_3\text{Sb}$ . This material is a semiconductor having a forbidden energy band about 1.5 eV wide. The photoelectric threshold energy is only slightly higher than this. Electrons excited from the occupied energy band by incident photons cannot assume energies lying in the forbidden band. They must remain in the conduction band shown in Fig. 5. Thus, even the slowest ones must retain energies only slightly less than that required for escape. The probability of photoemission is higher than for metals (or for semiconductors that have threshold energies greater than twice the width of the forbidden energy band).  $\text{Cs}_3\text{Sb}$  is sensitive over much of the visible range and can give very high yields, in excess of 0.2 electron per incident photon. It is widely used in practical phototubes. Related compounds can be made with enhanced photoelectric response in the red or ultraviolet regions of the spectrum.

**Alkali halides.** Three basically different kinds of photoemission are possible for alkali halides—intrinsic, extrinsic, and exciton-induced photoemission.

**Intrinsic photoemission.** This is characteristic of the ideally pure and perfect crystal. It is thus analogous to the emission already described for metals and semiconductors. It appears only for photon energies higher than the intrinsic threshold. For example, potassium iodide, KI, is an alkali halide having this intrinsic threshold in the far ultraviolet near 7 eV. Apparently, the width of the forbidden electron energy band in KI is only about 1 eV less than this. For the same reason that was mentioned

for the semiconductor  $\text{Cs}_3\text{Sb}$ , the photoelectric yields are high, in excess of 0.1 electron per incident photon, as shown by section C of the curve in Fig. 6.

**Extrinsic photoemission.** A second kind of emission occurs when a KI crystal contains imperfections in the form of negative ion vacancies (lattice sites from which negative iodine ions are missing). These vacancies can be filled by electrons. Color centers, which absorb visible light, are formed (see COLOR CENTERS). They may reach concentrations as high as  $10^{20}$  per  $\text{cm}^3$ . External photoelectrons may be ejected directly from these centers by photons. It is termed an extrinsic process since the light is absorbed by a crystal defect; it is also called direct ionization. The threshold energy for this process is about 2.5 eV. The yields can reach values of the order of  $10^{-4}$  electron per incident photon, as shown in section A of the curve in Fig. 6. The exact value of the yield depends on the concentration of color centers. Most of the incident radiation is lost because it is not intercepted by the centers, which present a limited cross-section to the incident photons.

**Exciton-induced photoemission.** When color centers are present, another photoelectric process takes place, in two stages. Potassium iodide has a sharp optical absorption band peaking at a photon energy of 5.6 eV. This is the first fundamental or intrinsic optical absorption band. Energy absorbed in this peak does not release free electric charges

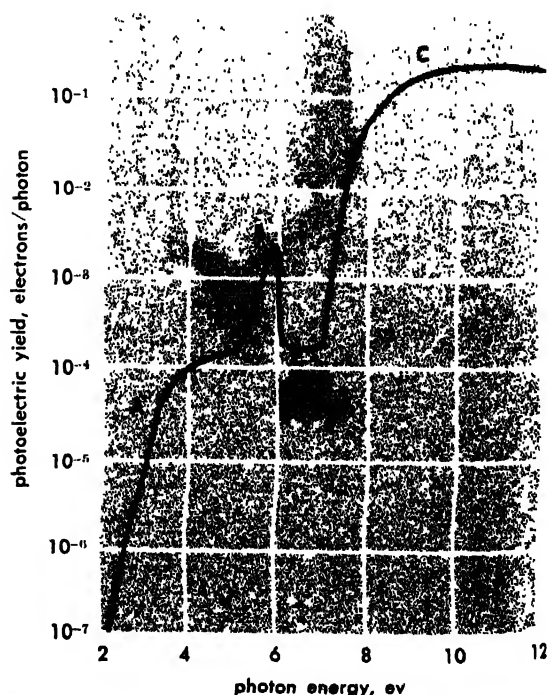


Fig. 6. Spectral distribution of photoelectric yield from potassium iodide containing color centers. Region A of the curve is due to direct ejection of photoelectrons from color centers; the peak B is due to exciton induced emission; C is due to intrinsic emission.

in the crystal. Rather, it leads to a kind of non-conducting excited state called an exciton state. The exciton can transfer enough energy to color centers to eject photoelectrons from the crystal. This two-stage process is termed exciton-induced photoemission. It appears in the peak B on the curve in Fig. 6. It is more efficient than direct ejection of photoelectrons from color centers. The entire crystal is capable of the primary photon absorption, and the energy can be transferred rather efficiently to color centers. Thus, the process avoids much of the loss in incident energy that arises from the limited cross section of color centers when they absorb photons directly. See EXCITON.

**Other compounds.** Other ionic crystals, such as barium oxide, behave much like the alkali halides. Direct ejection of photoelectrons from chemical impurities and from energy levels or defects localized at the crystal surface can be important. Besides these extrinsic processes, exciton-induced emission and intrinsic photoemission both occur.

Compounds such as zinc sulfide behave somewhat like germanium, but have higher intrinsic threshold energies, of the order of 7 ev. The photoelectric yields are comparatively low, as for germanium. Extrinsic processes such as direct ejection of electrons from chemical impurities (or defects) are sometimes detectable, but are usually weak.

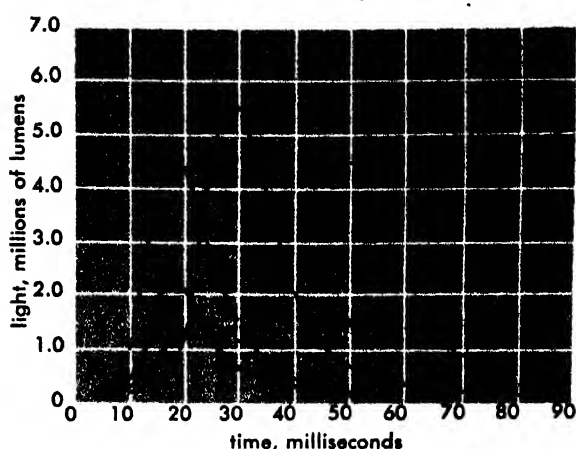
Certain complex photoemitters are made by letting cesium react with silver oxide to form cesium oxide and silver. They are valuable because they have threshold energies below 1 ev, and thus they are sensitive in the infrared. The photoelectrons appear to be directly ejected either from cesium adsorbed on the oxide surface or from discrete energy levels in the cesium oxide. The yields are about  $10^{-3}$  electron per incident photon. Intrinsic emission from cesium oxide (with yields above 0.01) occurs for photon energies above the intrinsic threshold at about 4 ev. [L.A.]

**Bibliography:** W. Shockley (ed.), *Imperfections in Nearly Perfect Crystals*, 1952; A. Sommer, *Photoelectric Tubes*, 2d ed., 1951; A. Van der Ziel, *Solid State Physical Electronics*, 1957; V. K. Zworykin and E. G. Ramberg, *Photoelectricity and Its Application*, 1949.

### Photoflash lamp

A combustion lamp that burns with a burst of high-intensity light, of short time duration and with definitely regulated time characteristics. The illustration shows the lumen-time characteristics of some typical lamps.

A photoflash lamp has a glass bulb filled with finely shredded aluminum foil in an atmosphere of oxygen. The foil is ignited by a low-voltage dry cell. The color temperatures in the various types range from 3800 to 6000°K. To protect against bursting, the outside and inside of the lamp has a plastic coating. Photoflash lamps are used for tak-



Lumen-time curves of some photoflash lamps.

ing ordinary photographs under poor ambient-light conditions, for high-speed photographs requiring fast shutter speed and high-intensity light, and for photographs requiring special lighting. See PHOTOGRAPHY. [J.O.K.]

### Photogrammetry

The practice of obtaining surveys by means of photography. Photographs may be taken on the ground or from an airplane. See SURVEYING.

Ground photographs normally are taken with the camera horizontal. Where position, orientation, and elevation of the camera and elevations of objects to be mapped are known, mapping can be performed with single photographs by reversing the procedure for constructing two-point perspective

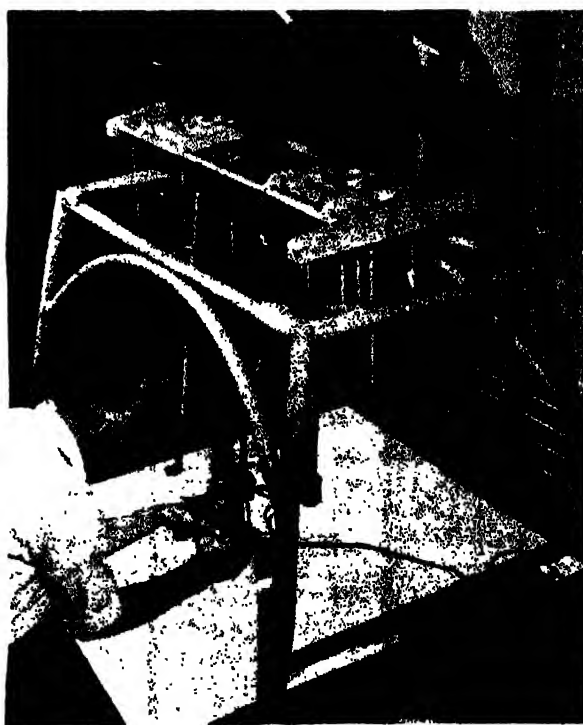


Fig. 1. Stereoplottting instrument. (Aero Service Corp.)



Fig. 2. Plotting table. (Lockwood, Kessler and Bartlett, Inc.)

drawings. Elevations of objects need not be measured where two photographs are taken from known points. Positions are plotted from graphic or analytic solutions of line-of-sight intersection problems.

Aerial photographs commonly are taken with the camera center vertical. Adjacent photos, taken at a prescribed altitude above a reference plane, are overlapped. The two images of the same terrain then can be viewed by using a stereoscopic technique so that one image is seen with one eye and the other image is seen superimposed with the other eye.

Although analytical methods exist for fixing the points measured on a stereoscopic image, direct mechanical solution of the problem is available. In one widely used procedure, glass positives of the two photos are placed in the projectors of a stereo-plotting instrument (Fig. 1). With the aid of identifiable position and elevation points (from ground surveys) the photos are oriented to the relative positions they had at the instants of exposure. Projections are aimed with the aid of space bars at a plotting-table top (white disk, Fig. 2), which can be raised or lowered. The image is in precise focus at the center of the disk only when the disk is at the correct scalar elevation for a given horizontal position. The center of the disk is indicated by a small spot of light called the floating dot. To trace a contour the operator sets its elevation on a dial integral with the table; he moves the table laterally until the elevation's focus is encountered by the floating dot, lowers the plotting pencil and follows the contour by keeping the image in focus at the floating dot. With special equipment, table movements can be recorded for electronic computer applications.

[R. H. DODDS]

## Photographic materials

The common sensitive materials of photography—plates, film, and papers. They consist of a support of glass, plastic sheet, or paper, respectively, coated with an emulsion, which in the usual instance is a suspension of silver halide crystals in gelatin and which provides the light-sensitive layer in which the picture will be formed.

**Supports.** The glass supports for plates are selected for optical clarity and flatness, and the thickness increases with the size of the plate, ranging usually from about  $\frac{1}{20}$  to  $\frac{1}{8}$  in. Film support, for many years mostly of flammable cellulose nitrate sheet, is now almost exclusively of the safety variety, consisting of a thin, flexible, transparent, optically uniform sheet of slow-burning material—cellulose acetate, and cellulose acetate propionate. When improved dimensional stability is required, film support consists of a variety of polymeric materials such as Vinylite, polystyrene, polycarbonate, and polyesters, particularly those related to substances derived from esters of terephthalic acid and ethylene glycol. Plasticizers are added to give flexibility in the case of the cellulose ester supports. Film support usually ranges from 0.00325 to 0.009 in. in thickness and is made in continuous rolls up to 60 in. wide and usually cut to about 2000 ft. Photographic paper is made from rag stock or mostly from wood pulp specially prepared to be free of chemical impurities, and has high wet strength. It is usually coated with a suspension of baryta (barium sulfate) in gelatin for high reflectance and may be calendered for high gloss.

Before the emulsion is coated on the support, in the case of plates and film, a "sub" or substratum is applied to ensure good adhesion of the gelatin layer (Fig. 1). In general, no sub is used with paper, although it might be needed in special cases, such as in water-repellent paper.

**Emulsions.** The emulsion consists basically of a suspension of silver halide crystals in gelatin, prepared by adding a solution of silver nitrate to a solution of halides in gelatin. The silver salts used in emulsions are the chlorides, bromides, and iodides. During manufacture, the emulsion is "ripened" and chemicals are added to control speed, image tone, contrast, and fog; to harden the

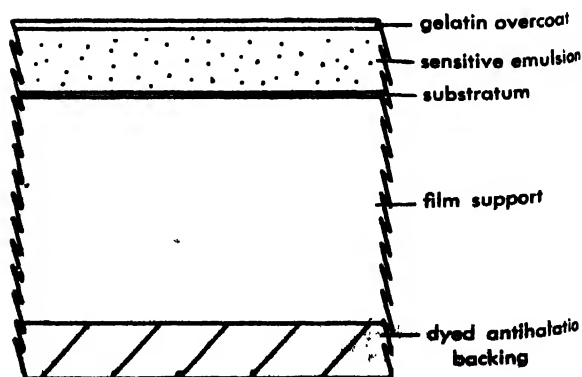


Fig. 1. Diagrammatic cross section of film.

gelatin; to assist in uniformity of coating; and to make the emulsion sensitive to desired wavelengths.

After being coated on the support, the emulsion is chilled to "set," and dried to a specific moisture content. Many films and some plates receive more than one coating, with individual layers being as thin as 0.0002 in. Most x-ray films are coated on both sides, some black-and-white films are double-coated on one side, and some color films have more than six coatings. Nonsensitized coatings may be applied over the surface of the emulsion to protect against abrasion during handling; roll film may have a gelatin coating on the back to reduce the tendency to curl by counteracting the effect of the gelatin emulsion on the front; and materials for making negatives usually have an antihalation coating on the back. Photographic paper emulsions are generally coated thinner than film emulsions and are more highly hardened.

**Spectral sensitivity.** The silver halides are normally sensitive only to the ultraviolet, violet, and blue but they can be made sensitive to longer wavelengths by adding to the emulsion special dyes, now usually of the carbocyanine and merocyanine class. This is known as optical sensitizing. In order to be able to reproduce tone values as seen by the eye, emulsions must respond to wavelengths to which the eye is sensitive, that is, approximately 4000-7000 Å, or from violet to red.

Nonsensitized emulsions are known as blue-sensitive, color-blind, or ordinary. Emulsions which have been treated with dyes to extend the sensitivity through the green are known as orthochromatic, and when the sensitivity is extended through the red they are known as panchromatic. On ordinary emulsions, blues are reproduced light and green and reds dark. Orthochromatic emulsions reproduce blues and greens as light and reds as dark, and they are used extensively in portraiture, commercial and industrial photography, and in the graphic arts. Panchromatic emulsions give reasonably good reproduction in black and white of the tone values of colored subjects and are particularly useful with incandescent light sources. The farthest extension of sensitizing by dyes is to about 13,000 Å, which is in the near infrared region. Ultraviolet sensitivity is limited by the strong optical absorption by gelatin below 2800 Å, but emulsions sensitive to much shorter wavelengths are prepared, with the gelatin greatly reduced.

**Photographic characteristics.** Many of the characteristics of sensitive materials, and the theory of the photographic process, latent image, and sensitometry, are discussed in another article (see PHOTOGRAPHY). All these characteristics are related to the combination of the emulsion, developer, and exposure. The structure of the developed image is of great importance for determining the quality of the photograph, in particular the definition or the ability to reproduce fine detail—the most important factors being graininess, resolving power, sharpness, and acutance. Graininess, the objective aspect of granularity, manifests itself as a nonhomogeneous, grainy appearance which may be

visible directly and is always visible under magnification. High-speed emulsions are generally grainier than slow emulsions, the graininess being dependent to a large extent on the nature of the development and the density of the image. The attempt is constantly made to increase speed without obtaining a correspondingly coarser grain. In prints, the graininess is higher as the density of the negatives from which they are made increases.

Resolving power, sharpness, and acutance depend on the turbidity (light scattering) and inherent contrast of the emulsion. Resolving power gives a measure of the ability to record fine detail and is usually expressed as the number of lines per millimeter which can just be separated visually. Resolving power values normally range up to about 150 lines/mm but may range from less than 50 to over 1000. They depend on the nature of the emulsion, subject contrast, density, and developer.

Sharpness refers to the ability of the emulsion to show a sharp line of demarcation between areas receiving different exposures. Usually such an edge is not sharp but graded to an extent depending upon the development conditions and the turbidity. Sharpness is related to the rate of change of density across such a boundary; the objective aspect worked out to describe this in terms of numbers is known as acutance.

**Photographic products.** Thousands of types and sizes of plate, film, and paper are available for a wide variety of applications. Generally, each field of use requires special properties. Amateur, professional, commercial, and industrial products for camera use may range from an exposure index of less than 10 to over 1000, from very fine grain and optimum sharpness to fairly coarse grain with corresponding loss in definition, and in a wide range of contrasts and spectral sensitivity, especially for scientific and industrial photography. Films of extremely high contrast and density are used in graphic reproduction in industry and the printing trade to give high-contrast line and halftone negatives. Blue-sensitive emulsions are used for copying and making duplicate negatives. Many types of x-ray film are made, some coated on both sides and for use with or without intensifying screens. Special multiple coatings are used for color photography. See PHOTOGRAPHY, COLOR.

A wide range of photographic papers available includes continuous-tone contact and enlarging papers in a range of contrasts for professional, amateur, and photofinishing printing, and a range of image tones and tinted paper support is available. A variety of papers having a range of speeds and contrasts is available for trace recording in oscillographs. High-contrast negative and direct positive papers are used for photographic reproduction of documents and engineering drawings. Papers for document reproduction by image-transfer systems are of two forms, (1) solvent transfer, in which the developed negative is placed in contact with a receiving sheet to which the undeveloped silver halide is transferred by a solvent and is reduced to a positive silver image, and (2) the system in which

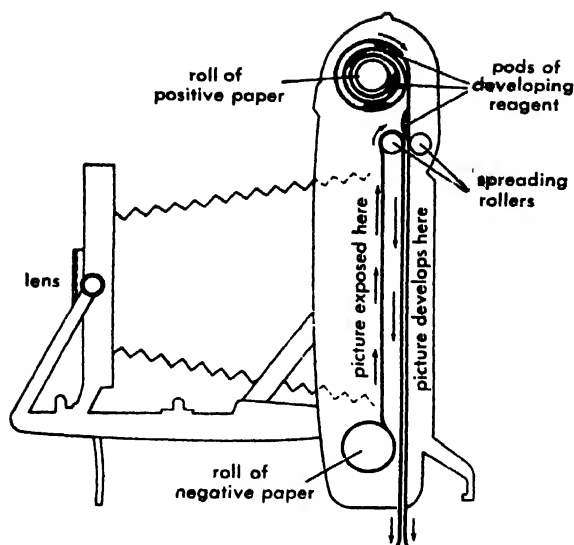


Fig. 2. Polaroid-Land camera. (Polaroid Corporation)

the processed negative is placed in contact with the receiving sheet to which the undeveloped emulsion is physically transferred and darkened.

**Polaroid-Land camera photography.** In this technique, the solvent-transfer process is used to give direct positive prints in the camera itself. The optical system of the camera (Fig. 2) is conventional and exposes a negative material in the back (see CAMERA). This is wound into contact with a separate roll of paper, using pressure rollers which burst a "pod" containing a thickened developer-silver halide solvent mixture which is spread as a thin layer between the two sheets. The negative is developed in the exposed sheet, and the undeveloped silver halide is dissolved and transferred to the receiving sheet where it is reduced to silver to form a positive image. This process is used for obtaining pictures in a minute. It is popular for amateur photography and has been adapted for press, commercial, and real-estate photography, aerial photography, reproduction of cathode-ray-tube images, and for other purposes where rapid production of a picture is desired.

**Storage of materials.** Films and papers may be damaged by high temperatures and especially high relative humidities. Color films are more seriously affected than black-and-white because the emulsion layers may be changed to different extents. Protection against high relative humidity is provided by special packaging, which is kept closed until the film is to be used. The package does not provide protection against heat. The lower the temperature, the better the film keeps, and for storage over several months a maximum of 45–55°F is desirable.

[W. CLARK]

**Bibliography:** See PHOTOGRAPHY.

## Photography

The process of forming visible images directly or indirectly by the action of light or other forms of radiation on sensitive surfaces. In the traditional sense photography utilizes the action of light to

bring about changes in silver halides which may be invisible, necessitating a developer to reveal the image, or which may be a directly visible darkening (print-out). Most photography is of the first kind and the function of the developer is to convert the exposed silver halide to silver. The bright parts of the subject give more exposure than the dark parts so that a negative results; that is, the brighter parts of the subject correspond to the darker parts of the reproduction. A positive, in which the relation between light and dark areas corresponds to that of the subject, is obtained when a negative is printed onto a sheet of similar material so that the negative tones are reversed. In the reversal process, direct production of the positive occurs if the developed negative silver is removed chemically and the remaining silver halide is then redeveloped; direct positive images can also be obtained directly by using special materials.

The common materials of photography consist of an emulsion of finely dispersed silver halide crystals (chloride, bromide, or iodide, depending on the purpose) in gelatin (Fig. 1) coated in a thin layer (usually less than  $\frac{1}{1000}$  in.) on glass, flexible transparent film, or paper (see PHOTOGRAPHIC MATERIALS). The most sensitive materials, used for negative-making, consist of silver bromide containing some silver iodide; the slow materials, used for prints, are usually of silver chloride; materials of intermediate sensitivity are of silver bromide or silver bromide and chloride.

After exposure of the emulsion-coated material in a camera or other exposing device, such as a spectrograph or recording instrument, the sheet is developed, fixed in a solution which dissolves the undeveloped silver halide, washed to remove the soluble salts, and dried. Printing from the negative is done by contact with or optical projection onto an emulsion-coated film or paper, and the same sequence of steps is followed as for the negative.

For about 100 years the results of practical photography were almost exclusively in black and white, but since the introduction of the Kodachrome process in 1935 a large and increasing percentage of photography has been done in color. In color photography, development is basically the same as in black-and-white photography except that the action is accompanied by formation of products which react to give dyes in addition to the silver, which is removed (see PHOTOGRAPHY, COLOR). Other photosensitive systems which are used in photography utilize diazo compounds, sensitive iron salts, photosensitive polymeric systems, bichromated colloids, bleachable dyes, photosensitive glass, and electrostatic, electrolytic, and photoconductivity effects. Gelatin is the common medium for the silver halide systems, but cellulose nitrate (collodion) has been used in photomechanical reproduction and polyvinyl acetate, albumen, and other colloid have been used. Glass is usually used as the support when flatness and rigidity are required. Transparent film support is of cellulose acetate, cellulose triacetate, cellulose acetate propionate, polystyrene, polycarbonate, or polyester material. The



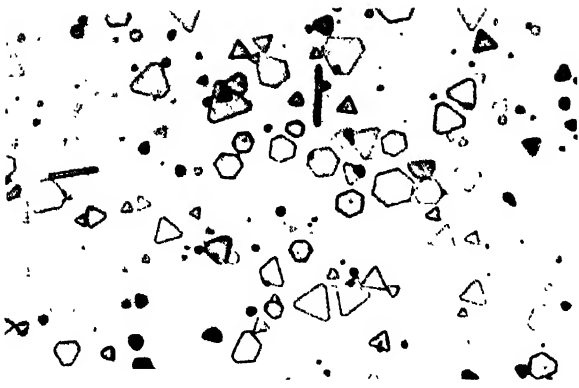


Fig. 1. Silver halide grains in a photographic emulsion, highly magnified.

nonemulsion-coated side of films and plates frequently carries a dyed coating to prevent broadening of the image by halation (light spreading) or a gelatin coating to counteract curl.

This article lists the major branches of photography, with subheadings on infrared photography, ultraviolet photography, high-speed photography, stereoscopic photography, document copying, photographic photometry, nuclear particle recording, microphotography (including microfilming); discusses the theory of the photographic process, including latent image formation, development, and after processes; and treats sensitometry, darkroom equipment, and contact and projection printers (enlargers). For related information, see AERIAL PHOTOGRAPHY; ASTRONOMICAL PHOTOGRAPHY; CAMERA; CINEMATOGRAPHY; LANTERN SLIDES; LENS, OPTICAL; MICRORADIOGRAPHY; MICROSCOPE, OPTICAL; OSCILLOSCOPE, CATHODE-RAY; PHOTOPLASTICITY; PHOTOGRAMMETRY; RADIOGRAPHY; SCHLIEREN PHOTOGRAPHY; SHOCK-WAVE DISPLAY; SPECTROGRAPHY; STROBOSCOPIC PHOTOGRAPHY; UNDERWATER PHOTOGRAPHY.

#### BRANCHES OF PHOTOGRAPHY

The main branches of photography are amateur, professional, commercial, educational, press, scientific and technical, and cinematography.

Amateur photography is the biggest branch and includes making photographs for record, for amusement, and for artistic purposes. Professional photography is concerned primarily with portraiture for commercial purposes. Commercial photography is mainly for advertising and industrial illustrative purposes. Educational photography is devoted to teaching and visual education. Press photography is for newspaper and magazine illustrations of topical events.

Photography is one of the most important tools in scientific and technical fields. It extends the range of vision, allowing records to be made of things which are invisible because they are associated with radiation to which the eye is insensitive or because they move too fast, or are too small, or too far away. Photographs are used as simple records which can be studied at leisure and measured and filed for reference or for security.

**Infrared photography.** Emulsions can be made to respond to radiations out to an upper wavelength limit of about 13,000 Å using special sensitizing dyes. Photographs can thus be made of subjects associated with radiation in the near infrared, such as spectra, stars, hot objects, and subjects which show selective transmission or reflection of near infrared radiation, especially in a manner different from visible radiation. Infrared photographs over long distances or from high altitudes show improved clarity of detail because the atmosphere may selectively transmit the near infrared and also because the contrast of ground objects may be higher as a result of their different reflectivities in the near infrared. Grass and foliage appear white because chlorophyll is transparent to the near infrared. Infrared photography has been used in camouflage detection because many green paints absorb infrared more strongly than does the foliage they may match visually. It has been used to record the distribution of temperature at the surface of heated objects, for photography in total darkness (the object being illuminated only by infrared), in criminology for deciphering altered or deteriorated documents and other objects, and for photographing textiles where dark dyes interfere with visual examination. It is used in medicine because the skin is somewhat transparent to infrared and the subcutaneous veins may be revealed for use in diagnosis. There are many applications in botany, paleontology, and technological fields, including the graphic arts. See INFRARED RADIATION.

**Ultraviolet photography.** This is used in the printing field, spectrography, and photomicrography and by mineralogists, police investigators, and examiners of questioned documents, and by museums and art galleries. The two methods of ultraviolet photography are (1) the fluorescence method, in which the subject is illuminated by ultraviolet and a filter is used on the camera to absorb the reflected ultraviolet and permit only the visible fluorescence to reach the film; and (2) the reflected ultraviolet method, in which an ultraviolet source is used and the camera is provided with a filter which permits only ultraviolet to reach the film. Ordinary, orthochromatic, and panchromatic films are used according to the purpose, and color photography can be done by the fluorescent-light method. See ULTRAVIOLET RADIATION.

**High-speed photography.** This serves a great variety of purposes in technological studies. Modern methods may be grouped as follows:

1. Single-exposure photography. The best mechanical shutters will not permit exposures shorter than about  $\frac{1}{1000}$  sec, although magneto-optical shutters, electrooptical shutters called Kerr cells (see KERR EFFECT), and electronic shutters have been used for very short exposures. Intense light flashes of very short duration are used in shadow photography and stroboscopic photography. In other reflected-light methods for single exposures, a common source is the gaseous-discharge lamp used with a normal camera; exposure times as short as  $\frac{1}{50,000}$  sec can be obtained.



2. High-speed series photographs are taken as motion pictures, the normal slow-motion cameras with intermittent motion giving up to 128 pictures/sec. For higher speeds (up to 5000/sec or more), continuous film movement is used with optical compensation for image movement, such as a rotating plane-parallel glass block or rings of prisms or lenses. Photographs made at these high frequencies, when projected at normal projection speeds, slow down the motion to the extent of the ratio of the taking and the projection speeds.

3. Sequences of short-duration photographs can be made at close intervals using flash bulbs, gaseous-discharge lamps, and groups of separate cameras operated in succession. Many special high-speed cameras giving high taking rates have been devised, including the Bell Laboratories' ribbon-frame camera and the F. E. Tuttle grid camera.

4. When the subject is self-luminous, of short duration, and moving rapidly, such as an explosion, a rapid sequence of photographs permits study of its development. Methods used include short individual exposures through a rapidly rotating shutter onto a stationary or moving film, falling in one plane at right angles to the path of the explosion or wrapped about a rotating drum. Rapidly rotating mirrors can be used in a moving film camera to give a series of separate photographs, or the film may be still and a single rotating mirror used. Continuous-trace photographs on moving film have been used to study explosions. High-speed single x-ray pictures have been taken by discharging a high potential of short duration through the x-ray tube.

**Stereoscopic photography.** This is done to simulate stereoscopic vision (see STEREOSCOPY). It presents to the two eyes individually two aspects of the subject made from slightly different viewpoints. Relief can be distinguished visually only over moderate distances, although the appearance of relief can be introduced in more distant objects by making photographs at greater distances apart. Stereoscopic photography can be done by using a single camera and making two separate photographs one after the other from viewpoints separated by the interocular or other appropriate distance, by displacing the lens of a single camera for the two exposures, or by rotating the object or the camera so as to give a pair of exposures on one film. These methods can be used only with relatively stationary subjects. Most stereoscopic photography is done by the simultaneous method in which two photographs are made at the same time, using two separate cameras, a stereoscopic camera (which is essentially two cameras in one body with matched optical systems and coupled focusing movements), or single cameras using beam splitters to give two photographs side by side on the film.

In aerial photography, the stereoscopic effect is achieved by making successive overlapping photographs along the line of flight; in terrestrial photogrammetry, photographs are made from each end of a selected base line and in stereoscopic radiography two x-ray photographs are made in rapid succession from separate viewpoints.

Stereoscopic photographs are viewed in equipment which presents the right-eye image to the right eye only and the left-eye image to the left eye only. This can be done by separate boxes, each provided with a lens; open-type viewers with a pair of shielded lenses, sometimes prismatic; cabinet viewers, with pairs of optics and devices for changing the stereo slides; grids or lenticular elements permitting each eye to see only its appropriate field; anaglyphs, in which two images are printed in ink in complementary colors and viewed through spectacles of similar colors such that each filter extinguishes one image; and the vectograph print, which consists of a reflecting support on which the stereoscopic pairs are placed one over the other in plastic polarizing layers with polarization planes at right angles. To view, polarizing spectacles are used, the eyepieces being arranged so that each eye sees only the appropriate image. See VECTOGRAPH.

**Document copying.** Photography is used for reproducing documents of all kinds because (1) errors are not introduced in the copying process, (2) it effectively extends the life of perishable records, (3) it offers security against loss and disaster by permitting storage of multiple copies, and (4) it can provide for retrieving and disseminating the information in them rapidly.

Documents are reproduced to essentially the same size as the original by special copying papers and apparatus, and in reduced size, usually on film, in reducing cameras (microfilming). Prints from the films, usually 16-, 35-, 70-, and 102-mm, can be made by contact onto silver- or diazo-sensitized films, sometimes for insertion into window cards, or by conventional photographic enlarging or by xerography (an electrostatic process) to give reproductions in any scale.

Paper copies of documents to scale or at moderate reduction are made by contact printing (printing through or by reflex copying) or in cameras, and may be negatives or positives. In the case of negative copies (for example, Photostat prints) optical reversal is used on the camera to give correct orientation. In contact printing a laterally reversed negative is usually obtained, from which positive prints may be made by printing through. Positive copies are obtained directly by using special direct-positive papers. Image transfer processes are used to give positive prints on receiving sheets by transfer from reflex-printed papers (such as Kodak, Verifax, Agfa Copyrapid, Gevaert Geva-copy). Systems other than silver halide systems make copies by using thermography (Thermofax) or electrostatic (xerography, Electrofax) and electrolytic processes. See PHOTOCOPYING PROCESSES.

**Photographic photometry.** The intensity of radiation, or the spectral distribution of intensity, can be measured by photography. The radiation whose intensity is to be measured is compared with that from a standard source by matching the photographic densities produced by both. The method is capable of high precision if the characteristics of photographic materials are accurately known and the results are interpreted intelligently. It is quite

unreliable to attempt to compute energy from measurement of a single density. In practice, a single exposure is made to the unknown radiation, and an adjacent series of exposures (closely adjacent on the same plate or film) is made to the standard source, the exposure in each step of the series being accurately known. The density on the developed plate or film for the standard which matches that of the unknown is selected and the intensity which gave it is also that which gave the unknown. The conditions to be fulfilled are strict and have been defined by L. A. Jones (1937) and G. R. Harrison (1929). See **PHOTOMETRY**.

**Nuclear-particle recording.** Charged atomic particles give records on photographic emulsions, and the method of recording provides an important adjunct to such detectors as the ionization chamber, the Geiger counter, and the Wilson cloud chamber. The first studies were with  $\alpha$ -particles, which produce a track of silver grains in the developed emulsion. Protons were later found to produce tracks, with grains of different spacing. Cosmic rays give nuclear disintegrations on collision with the atoms in the emulsion; starlike patterns result, consisting of tracks made by the particles from the atomic nuclei. Tracks of electrons and other charged particles can be recorded on special thick emulsions having high silver bromide content, small grains, minimum fog, and appropriate speed. The grains are made developable by ionizations within them caused by impact of the particles. See **COSMIC RAYS**, **PARTICLE DETECTOR**.

**Microphotography.** The process of making photographs on a greatly reduced scale is called microphotography. Microfilming is the special technique of copying documents to reduced size on film, usually 16 and 35-mm film, but 70- and 102-mm widths are used as well as sheet film. The negatives, or contact positive films or paper prints made from them, can be read in enlarging readers,

or enlargements onto paper may be used to provide copies. Special cameras are used, provided with magazines to hold 100 ft or more of film and sometimes having variable magnification, automatic focusing, and high-definition lenses. Some cameras are in the form of desks into which the documents are fed rapidly and photographed and moved through the machine automatically. Reductions may range from 8 to 40, but in extreme cases, by using lenses and films of exceptional definition, as in the Minicard system, a reduction of 60 times has been used. Special projection equipment made for reading the films may give an enlargement onto a table or a diffusing rear-projection screen, with provision for rapidly winding the film and framing any desired page. In reader-printers, the image viewed on the screen may be used at will to give a paper print.

Films may be mounted individually in apertures in cards for use in sorting machines. Strips of film may be assembled in sheaths in sheet form, or as "microfiches" in which a great many separate pictures are printed onto a sheet of film.

#### THEORY OF THE PHOTOGRAPHIC PROCESS

The normal photographic image consists of a large number of small grains of silver. They are the end product of exposure and development from the original silver halide crystals of the emulsion, which range up to a few microns in size (see Fig 2). The size and distribution of size of the crystals are determined by the way the emulsions are made; the sizes are closely related to the photographic properties. In general, the high-speed negative-type emulsions of moderate contrast have crystals in a wide range of sizes, whereas the emulsions with low speed and high contrast have small crystals fairly uniform in size.

When an emulsion is exposed to light, an invisible change called the latent image is produced.



Fig. 2. (a) Developed silver grains of a photographic emulsion, (b) Original grains of silver halide from

which the silver grains in (a) were developed; both highly magnified.

The effect of exposure is made visible by development in a chemical reducing solution which converts to silver the crystals unaffected by the exposure, leaving unreduced those not affected. The darkening (density) is determined by the amount of exposure and the extent of development and depends on the number of silver grains developed in a particular area.

With very high exposures, density may form directly without development as a result of direct photolysis of the silver halide giving silver. This is known as the print-out effect. Its use is confined mostly to making proof prints in portraiture and in certain direct-trace-recording instruments.

Development of exposed crystals starts at isolated points on the surface. These appear to be associated with points having special sensitivity, indicating that the latent image is concentrated at specific points. These so-called sensitivity centers appear to be associated with specks of silver and silver sulfide in the crystal.

**Latent image.** The term latent image refers to the change occurring in the individual crystals of photographic emulsions whereby they become developable on exposure to light. Once development has been initiated, it continues in an individual grain until effectively all the silver halide is converted to silver.

The concentration speck theory of S. E. Sheppard, A. P. H. Trivelli, and R. P. Loveland forms a basis for modern latent image theory and suggests that the light energy absorbed by the crystal is concentrated in the vicinity of the silver-silver sulfide specks, where it liberates silver from the silver halide. Most evidence indicates that the latent image is mostly silver and that the silver speck can initiate development of the crystal when it has grown to a certain size. The present theory of the mechanism of latent image formation suggests that on exposure the electrical conductivity of silver bromide is increased by electrons becoming available. J. H. Webb suggested that these electrons could move freely through the crystal and become trapped at the sensitivity specks. Also, conductivity of silver bromide is caused by movement of silver ions. R. W. Gurney and N. F. Mott visualized a combination of these two processes to explain the formation of the latent image as follows.

On exposure, the first reaction involves liberation of electrons when energy is absorbed by the crystal. These electrons are able to move freely in the crystal until they impinge on the sensitivity specks, where they are trapped and build up a negative charge. This charge attracts the positively charged free silver ions which wander to the specks, where they are neutralized by the electrons to give neutral silver atoms, as a result of which the specks grow until they are big enough to act as development nuclei. The Gurney-Mott theory is the best basis for the explanation of latent image formation at the present time. For additional information on the Gurney-Mott theory, see PHOTOLYSIS (PHOTOCHEMISTRY). See also PHOTOCONDUCTIVITY.

Latent images are formed not only on surfaces

of the crystals but also in their interiors. In the latter case the latent image is presumably associated with localized structural imperfections in the crystals. The internal latent image is frequently formed by very-high-intensity exposures of short duration.

**Development.** Development is of two kinds physical and chemical. Both physical and chemical developers contain chemical reducing agents, but a physical developer also contains silver compounds in solution (directly added or derived from the silver halide by a solvent in the developer) and works by depositing silver on the latent image. Physical development as such is little used, although it usually plays some part in chemical development. A chemical developer contains no silver and is basically a source of reducing agents which will distinguish between exposed and unexposed silver halide and convert the exposed halide to silver. Developers in general use are compounded from organic reducing compounds, an alkali to give desired activity, sodium sulfite which acts as a preservative, and potassium bromide or other compounds used as antifoggants (fog is the term used to indicate the development of unexposed crystals; it is usually desirable to suppress fog).

Most developing agents used in normal practice are phenols or amines, and a classical rule which still applies, although not exclusively, states that developers must contain at least two hydroxyl groups or two amino groups, or one hydroxyl and one amino group attached ortho or para to each other on a benzene nucleus. Some developing agents do not follow this rule but among the common developers which do are hydroquinone, monomethylparamino-phenol (Elon, Metol), pyrogallol (1,2,3-hydroxybenzene), Amidol (2,4-diaminophenol), and *p*-phenylenediamine. In 1951 Ilford Ltd produced another kind of developer, Phenidon (1-phenyl-3-pyrazolidone), which can replace a great part of Metol in many metol-hydroquinone developers. Metol and hydroquinone are frequently used together.

Alkalies generally used are sodium carbonate sodium hydroxide, and sodium metaborate (Kodalk). Sulfite in a developer acts by lowering the tendency for oxidation by the air. Oxidation products of developers have an undesirable influence on the course of development and may result in stain.

Developing agents and formulas are selected for use with specific emulsions and purposes. Modern color photography relies mainly on paraphenylene diamine derivatives. So-called fine-grain developers are made to reduce the apparent graininess of negatives. They generally contain the conventional components but are adjusted to low activity and contain a solvent for silver bromide. One fine-grain developer is based on *p*-phenylenediamine, which itself is a silver halide solvent. Some developers are compounded for hardening the gelatin where development occurs, the unhardened area being washed out to give relief images for photo-mechanical reproduction and imbibition color printing. In certain materials, such as the Verifax matrix material, the hardening developing agent is

included in the emulsion, development being initiated by applying an alkali called an activator. For special purposes monobaths, consisting of a developer containing a fixing agent, are used.

When an exposed film is developed, there is usually a period during which no visible effect appears; after this the density increases rapidly at first and then more slowly, eventually reaching a maximum. In the simplest case, the relation between density and development time is

$$D = D_{\infty}(1 - e^{-kt})$$

in which  $D$  is the density attained in time  $t$ ,  $D_{\infty}$  is the maximum developable density, and  $k$  is a constant called the velocity constant of development.

**After processes.** These include fixing, washing, drying, reduction, intensification, and toning.

**Fixing.** After the image is developed, the unchanged halide is removed, usually in water solutions of sodium or ammonium thiosulfate (known respectively as hypo and ammonium hypo). This removal of unchanged halide is called fixing. Prior to fixing, a dilute acid stop bath is often used to neutralize the alkali carried over from the developer. Alternatively, an acid fixing bath containing thiosulfate, sulfite, and acetic acid is commonly used, and hardeners may be added to prevent softening of the gelatin. The rate of fixing depends largely on the concentration of the fixing agent and the temperature. In the case of sodium thiosulfate, the rate of fixing is most rapid at 20–40% concentration and at a temperature of 60–75°F. In so-called “stabilization” the undeveloped silver salts are converted into more or less stable complexes and are not washed out.

**Washing.** Negatives and prints are washed in water after fixing to remove the soluble silver halide-fixing agent complexes, which might render the photographs unstable on keeping or might cause stain. The rate of removal of the compounds by washing is exponential, and is increased by raising the temperature and increasing the agitation. The rate can be accelerated by neutral salt solutions known as hypo clearing aids, thus reducing the washing time. Remaining traces of thiosulfate may be removed by chemical treatment which converts the thiosulfate to sulfate that does not affect the image on storage.

**Drying.** After washing, the materials must be dried, preferably in moving warm air. In the case of paper prints, drying is frequently done on heated metal drums which may also give gloss to the prints if they are dried with their emulsion side to the metal surface.

**Reduction and intensification.** Reduction refers to methods of decreasing the density of images by chemically dissolving part of the silver by using oxidizers. According to the composition, oxidizers may remove equal amounts of silver from all densities, or remove silver in proportion to the amount present, or remove more silver from the higher densities than from the lower.

**Intensification** refers to methods for increasing density of an image, usually by deposition of

silver, mercury, or other compound, the composition being selected according to the nature of the intensification required.

**Toning.** Photographs are normally designed to be reasonably neutral in color. By special treatments known as toning, the color can be modified; for instance, by additions to the developer, by the selection of special developers, by toning solutions which convert the silver image into such compounds as silver sulfide, or by precipitating colored metallic salts with the silver image. Dye images are obtained by the methods of color development.

## SENSITOMETRY

Strictly speaking, sensitometry refers to the measurement of the sensitivity or response to light of photographic materials, but in practice it includes a variety of other factors which determine the properties of the final image. The simplest method of determining sensitivity is to give a graded series of exposures and to select the exposure required to give the lowest visible density. Modern sensitometry depends on plotting curves showing the relation between the logarithm of the exposure  $E$  and the density of the silver image in relation to the development, spectral quality of the light source, and other factors. Density  $D$  is defined as the logarithm of the opacity  $O$ , which is in turn defined as the reciprocal of the transparency  $T$ . If light of intensity  $I$  falls on a negative, and intensity  $I'$  is transmitted,

$$T = I'/I \quad O = 1/T = I/I' \quad D = \log I/I'$$

The relationship between density and the logarithm of exposure is shown by the Hurter and Driffield characteristic curve, which has three fairly well defined proportions, the greater part of which in most cases approximates the straight line  $B$  to  $C$  in Fig. 3. In this region the exposures are directly proportional to the brightness values in the subject. In the lower and upper portions of the curve this proportionality does not apply; these portions are the underexposure region ( $A$ – $B$ ) and the overexposure region ( $C$ – $D$ ), respectively.

The characteristic curve is used to determine the sensitivity or speed of a material, the con-

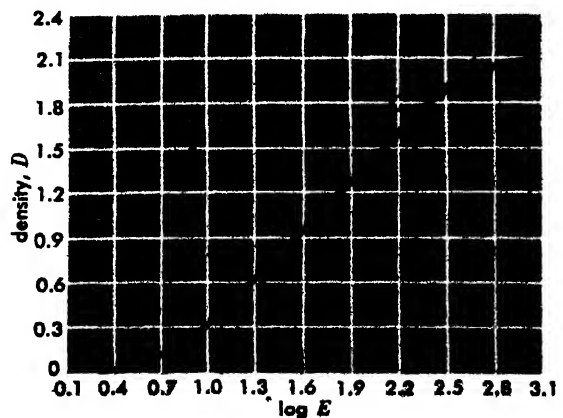


Fig. 3. Characteristic curve.

trast, the exposure latitude, and the tone reproduction. It is obtained under controlled conditions using a light source of known intensity and spectral characteristics, a modulator which gives a series of graded exposures of known values, development under precise conditions, and an accurate means of measuring the densities.

The internationally adopted light source is a tungsten-filament electric lamp operated at a color temperature of 2360°K (that is, 2360°K is the temperature of a black body whose radiation has the same energy distribution as that from the tungsten surface), combined with a filter to give spectral quality approximating that of mean noon sunlight in Washington, D.C., namely, 5400°K. (The color temperature of the sun as filtered by the atmosphere when the sun is on the meridian at the latitude of Washington, D.C. is 5400°K; see COLOR TEMPERATURE; HEAT RADIATION.) The light source and exposure modulator are combined in a sensitometer, which gives a series of exposures increasing stepwise or in a continuous manner. Sensitometers are either intensity-scale or time-scale instruments, depending on whether the steps of exposure are made at a constant time and varying intensity or at constant intensity and varying time. The exposure may also be intermittent, that is, a series of short times adding up to the desired total time. Continuous exposures are preferred to avoid the intermittency effect. (Below certain critical high frequencies the photographic material does not add up separate short exposures arithmetically.) The best sensitometers are of the continuous-exposure intensity-scale variety, and the exposure time used should approximate that which would be given the material in practice.

Development is carried out in a specified formula or in the developer recommended for use with the material under test for the desired time or series of times at a standard temperature of 68° or 75°F and with agitation which will ensure uniform development and reproducibility.

**Densitometers.** Photographic density is measured in a densitometer. Early densitometers were

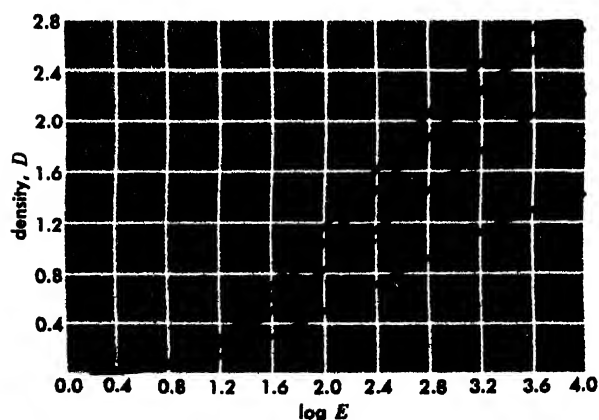


Fig. 4. Characteristic curves for development times increasing in the order 1, 2, 3, 4; abscissa scale is in meter-candle seconds.

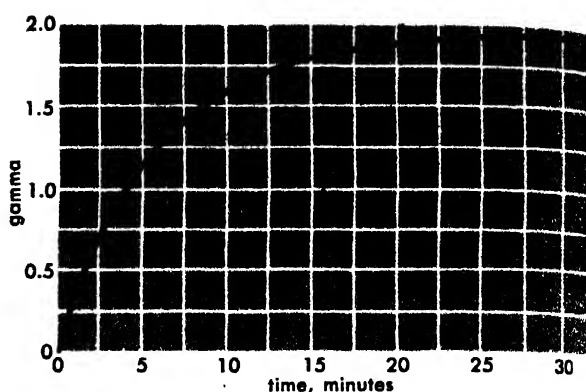


Fig. 5. Gamma versus time of development curve.

based on visual photometers. Usually two beams are taken from a standard light source and brought together in a photometer head, one beam passing through the density to be measured and the other through a device for modulating the intensity that the two fields can be matched. Visual instruments are still used, but largely as primary reference standards. Most densitometers today are the photoelectric type, the intensity-modulating device usually being a standard density which has been calibrated on another instrument, or variable apertures.

The simplest photoelectric transmission densitometers are of the deflection type consisting of a light source, a barrier-layer cell, and a microammeter, the density being obtained from the relative meter deflections with and without the sample in place. Other photographic densitometers employ the null system, in which the current resulting from the transmission of the test sample is balanced by an equivalent current derived from another photocell or from the same cell illuminated by the light beam through a calibrated modulator in alternation with the light beam from the sample. The trend in practice is to use another null method in which the modulator is placed in the same beam as the test sample and so controlled that the total transmittance of the sample and the modulator is kept constant by balancing against a constant comparison beam. Many densitometers are equipped with devices for automatically plotting the characteristic curves. The densitometry of colored images presents special problems and reference should be made to the publications of C. E. K. Mees, and R. M. Evans, W. T. Hanson, Jr., and W. L. Brewer listed in the bibliography.

When light passes through a negative, some is specularly transmitted and some is scattered. If all the transmitted light is used in densitometry, the result is called the diffuse density. If only the light passing directly through is measured, the density is called specular. The diffuse density is higher and is related to specular density by a relationship known as Callier's  $Q$  factor.

In the case of paper prints, the density is measured by reflection. The reflection density is  $D_R = \log (1/R)$  where  $R$  is the ratio of light reflected by

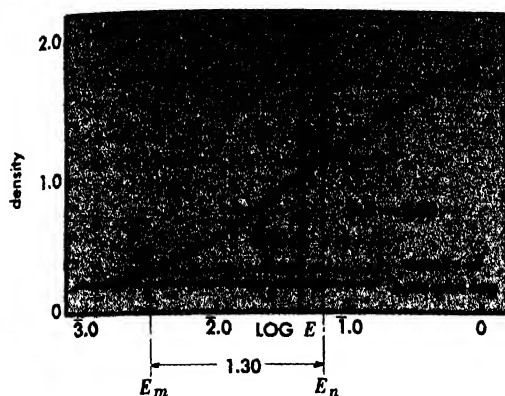


Fig. 6. Derivation of speed according to the 1960 American Standard Method.

the paper to that of the image on it.

If the straight line of the characteristic curve  $BC$  is extended to meet the exposure axis at  $i$  (called the inertia) on Fig. 3, the angle  $\alpha$  is used to express contrast.

The tangent of  $\alpha$ , that is, the slope of the line  $BC$ , is known as  $\gamma$ . As the time of development increases, the slope of the curve increases, eventually reaching a maximum (Figs. 4 and 5). The projection of the straight line  $BC$  of Fig. 3 onto the exposure axis at  $bc$  is the latitude, and the distance  $ad$  represents the total scale, that is, the whole exposure range in which brightness differences in the subject can be represented.

**Speed.** The sensitivity or speed of a material is derived from the characteristic curve. The earlier Hurter and Driffield (H. and D.) speeds were obtained by dividing 34 by the exposure corresponding to the inertia. Some methods (Eder-Hecht, Scheiner) were based on the exposure required to give a density which was just perceptible. The idea was developed by R. Luther and later by L. A. Jones that speed systems should be based on a gradient in the toe of the curve, and specifically by Jones on the least exposure to give negatives from which prints of excellent quality could be made. British and American standard speeds were based on the exposure  $E$  in the toe of the curve where the gradient is 0.3 of the average gradient over a log exposure range of 1.5. The relationship between speed  $S$  and exposure  $E$  is  $S = 1/E$ . To determine camera settings for pictorial photography, exposure indexes, which were equal to  $1/4E$ , were used instead of speed.

A safety factor was incorporated in the exposure index for black-and-white negative materials, and when these ratings were used with most exposure meters a factor of  $2\frac{1}{2}$  was effective. It was later realized that in practice the extra high exposure resulting from these indexes could be a disadvantage, and strong arguments were made to reduce or eliminate the safety factor. It was also found that a fixed density-above-fog criterion could correlate well with fractional gradient speed if development was to a fixed average gradient.

In 1960 a revised American standard was issued, and the method for determining speed according to it is shown in Fig. 6. Point  $M$  is at a density of 0.10 above fog-plus-base density. Point  $N$  lies 1.3 log  $E$  units from  $M$  in the direction of greater exposure. Developing time in the standard developer must be chosen so that  $N$  lies at a density 0.8 above the density at  $M$ . Then, the exposure  $E_m$  (in mcs) corresponding to  $M$  represents the parameter from which ASA speed is computed.  $S = 0.8/E_m$ .

Characteristics of printing papers are determined on the same general lines as negative materials, important factors being exposure scale, density scale, and useful maximum density and speed, the latter being based on a gradient in the shoulder of the curve.

The density resulting from an exposure is usually not independent of the absolute values of intensity and time, an effect known as failure of the reciprocity law.

**Tone reproduction.** This refers to the relation between the luminance and luminance differences in the subject and the density and density differences in the photograph. It has objective and subjective aspects and has been thoroughly worked out. Reference should be made to the work of L. A. Jones and C. E. K. Mees in the bibliography.

#### PHOTOGRAPHIC APPARATUS

The camera is the basic instrument of photography, and is discussed separately (see CAMERA). Other important apparatus of photography includes the means for lighting the subject, equipment for printing by contact or projection and for viewing transparencies, devices for handling films, plates, and paper in the various stages of processing, and means for viewing, storing, and retrieving photographs.

**Darkroom equipment.** The darkroom may range from total darkness to room light with safelights colored according to the materials handled. When loading and unloading of film and plates in holders or cameras and processing of films, plates, and papers must be done in the open, these operations are done in the darkroom. Basic equipment consists of safelights, which are lamp houses with filters to illuminate the room with light of a color which will not fog the sensitive material in a reasonable time and which give maximum visibility consistent with this safety; benches and sinks with running water and drains; tanks and hangers for processing plates and sheet films vertically; tanks for roll films which are usually cylindrical and light-tight and contain a reel onto which the film is wound in a spiral with spaced convolutions to permit access of the solutions; spiral reels onto which film can be wound for processing in a tank or tray; flat trays for sheet film, plates, and paper sheets and occasionally short roll-films; thermometers for determining temperatures of solutions or mixing valves to give water of desired temperature for controlling solutions in tanks;



special multiple tank units for color photography; tanks, trays, or special washers for washing negatives and prints; dryers, ranging from simple clips for hanging negatives in the open air to cabinets having forced warm air, special dryers for prints including simple blotting-paper sheets, and heated flat, drum, or belt dryers including some which ferrotype (that is, gloss) the prints by drying them in contact with a polished surface; clocks and preset timers; printers and enlargers; print trimmers; and for specialized work, densitometers, printing exposure meters, and focusing devices.

For professional processing of negatives and printing and processing of prints on a large scale, (for example, in the photofinishing industry), continuous film and roll-paper-processing machines and automatic or semiautomatic printers for contact printing or enlarging onto sheets or rolls of paper are used. Continuous machines are made for processing x-ray negatives in sheet form on a large scale, and many continuous-processing machines have been designed for special purposes, such as aerial photography, microfilming, scientific recording, and motion-picture film.

**Contact and projection printers.** Negatives provide the primary records of photography in the black-and-white field and to an increasing extent in color, and in many cases, such as the document reproduction field, they may be more useful than the positives. In general, however, positive prints are required, and such prints are made from negatives in contact printers and projection printers (or enlargers).

**Contact printers.** These are boxes containing lamps, a glass top on which a negative is placed, over which and in contact with which is put a sheet of printing paper and a lid or platen which is provided with springs, air bellows, felt, spongy material, or other means of applying uniform pressure to press the negative and paper into good contact. Vacuum or air pressure may also be used to provide contact. Contact printers print on single sheets or rolls of paper from single negatives or rolls of negatives. Exposure timers and photoelectric cells for automatically controlling the exposure may be incorporated. Positive film transparencies may be printed from negatives, particularly lantern slides and motion-picture film in long rolls.

Reflex printing is done by contact, by exposing through the base side of the photographic paper, the sensitive side of which is in contact with the original document. In bireflex printing the exposure is also made through the sensitized paper, but its base side is in contact with the document, the base being translucent. In contact printers for thermographic processes the reflex principle is used, the source of radiation in the printer being mainly heat.

**Projection printers (enlargers).** These are optical projectors consisting of a lamp house and light source, a holder for a negative, a condenser or diffusing sheet, a projection lens designed to give optimum quality at relatively low magnifications,

means for focusing for the desired magnification, sometimes automatically coupled to ensure good focus at all magnifications (autofocus enlargers), a support for the optical components, and a board or easel to carry the sensitive paper. Timers, manually or photocell operated, may be incorporated to control exposure time. Projection printers may expose individual paper sheets or rolls of paper from single negatives or rolls of negatives.

In some printers (for example, LogEtronics printers) the negative is scanned by a spot of light which at the same time exposes the printing paper, the exposure at any point being controlled by the response of a photo-cell to the scanning beam.

[W. CLARK]

**Bibliography:** H. Baines, *The Science of Photography*, 1958; T. T. Baker, *Photographic Emulsion Technique*, 2d ed., 1948; A. Boni, *Photographic Literature*, 1962; W. Clark, *Photography by Infrared*, 2d ed., 1946; L. P. Clerc, *Photography, Theory and Practice*, 3d ed., 1954; G. T. Eaton, *Photo Chemistry in Black-and-white and Color Photography*, 1957; R. M. Evans, W. T. Hanson, Jr., and W. L. Brewer, *Principles of Color Photography*, 1953; Focal Press, *The Focal Encyclopedia of Photography*, 1965; P. Glafkides, *Photographic Chemistry*, vol. 1, 1958, vol. 2, 1960; K. Henney and B. Dudley (eds.), *Handbook of Photography*, 1939; R. W. G. Hunt, *The Reproduction of Colour*, 1957; T. H. James and C. C. Higgins, *Fundamentals of Photographic Theory*, 1948; G. A. Jones, *High-speed Photography, Its Principles and Applications*, 1952; A. W. Judge, *Stereoscopic Photography*, 3d ed., 1950; R. Kingslake, *Lenses in Photography*, 1951; E. K. Mees, *The Theory of the Photographic Process*, rev. ed., 1954; W. D. Morgan (ed.), *The Encyclopedia of Photography*, 20 vols., 1963-1964; C. B. Neblette, *Photography, Its Materials and Processes*, 6th ed., 1962; S. E. Sheppard and C. E. K. Mees, *Investigations on the Theory of the Photographic Process*, 1907; D. A. Spencer (ed.), *Progress in Photography*, 3 vols., 1940-1958; G. W. W. Stevens, *Microphotography*, 1957.

## Photography, color

A large proportion of photography is done in color. The camera exposure may result in black-and-white separation negatives, a positive color transparency, or a negative color transparency; any of these may be used to produce prints on film or paper or other white support by a variety of color printing processes. See PHOTOGRAPHIC MATERIALS; PHOTOGRAPHY.

The two basic processes of color photography are the additive and subtractive systems. Since 1935 the subtractive processes have been used almost exclusively. The first step in both cases is to make photographs of the subject by its blue, green, and red components (hence the expression three-color photography), dividing the spectrum into three roughly equal parts by means of color filters or selective spectral sensitization of adjacent emul-



sion layers. Two-color processes are used to a limited extent, dividing the color into two components, blue-green and orange, but the color rendering is greatly inferior to that of the three-color method.

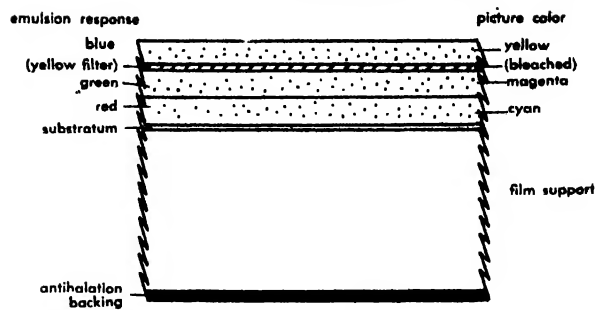
**Additive processes.** In these, the colors are produced by adding blue, green, and red in the amounts present in the original subject. There are three main systems: triple projection or viewing, the screen-unit system, and the lenticular film process. In the first, three negatives are made in succession in a single camera, or simultaneously in a one-shot camera, or on a tripack, using panchromatic film (film sensitive to light of all colors), and black-and-white positive transparencies made from these are projected or viewed directly in superposition, each positive being illuminated by the color by which its negative was made. This method is now little used.

In the screen-unit process, typified by the Lumière Autochrome plate and Dufaycolor film, the plate or film is covered with a layer of minute blue, green, and red elements, over which a panchromatic emulsion is coated. Exposure through the colored elements (screen units) and development by the reversal process gives a positive black-and-white image broken up into elements corresponding to the screen units and of density determined by the color of the original. Direct viewing of the combination gives a color reproduction.

The lenticular film process (Berthon, Keller-Dorian, Eastman Kodak's early Kodacolor process) used a film embossed on the back with minute lenses, which in the camera faced the lens, which carried a three-color banded filter. Development was by reversal and the photograph was broken up into minute elements each of which was an image of the filter, and of density corresponding to the color of the subject at that point. Projection through a similar banded filter gave a color reproduction. The lenticular system has been used for monitoring color television.

**Subtractive processes.** In these techniques, the subject is photographed by blue, green, and red light, but from the negatives, positives are made in dyes or pigments in colors complementary to those by which the corresponding negatives were made, in this case respectively yellow, magenta, and cyan, and the three colored positives are viewed in superposition on transparent film or white paper. The three separation negatives are made in the ways mentioned for additive processes or as monopacks or integral tripacks in which they are not separated but converted directly into multilayer color photographs.

The monopacks provide the basis for most color photography. Essentially they consist of three emulsions coated one above another and sensitive respectively to blue, green, and red light, the blue-sensitive layer being on top with a yellow filter layer between it and the green-sensitive layer to prevent penetration of blue light to the lower layers. The three color images in the separate layers



Cross section of Kodachrome film.

are formed by using developers which give oxidation products which combine (couple) with dye-forming components (couplers) in the developer or the emulsions, the couplers being chosen to give a yellow dye in the blue-sensitive layer, magenta dye in the green, and cyan dye in the red.

In Kodachrome film, the first successful monopack (1935) used for making positive color transparencies on film, the exposed film is developed to give a black-and-white negative in each layer. It is then exposed through the base to red light, which exposes the undeveloped silver halide in the red-sensitive bottom layer; development in a developer containing a cyan coupler gives a cyan positive image. Exposure to blue light from the front and development with a yellow coupler present gives a yellow dye positive in the top layer. The middle layer is made developable to give a magenta dye image, the silver in all three layers is bleached out, and a three-color transparency remains.

In the other monopack method of making positive color transparencies, the dye-forming couplers are incorporated in the emulsions, thus permitting simultaneous production of the three dye images in a single development step. In one form the couplers are attached to heavy molecules which prevent them from diffusing, so that each dye is formed in its appropriate layer. In the other form the couplers are made stationary by dissolving them in oily liquids which are finely dispersed in the emulsions. In both processes, to obtain positive color transparencies, a first development gives black-and-white negatives in all three layers, color development gives color positives from the residual silver halide in the layers, and bleaching the silver from all layers leaves a color transparency.

Negative color films (Kodacolor, Ektacolor, Agfacolor Negative, Gevacolor) are obtained from monopacks similar to those described but developed directly in a color developer, followed by bleaching the silver, the result being a color transparency in complementary colors. In the case of Kodacolor and Ektacolor negative films, the couplers used are actually colored red and yellow (apart from the developed color) and these colors are destroyed in proportion to the amount of dye image developed, so that the negative dye image is associated with a positive image in red and yellow. This colored mask offsets lack of purity of the negative dyes and gives prints of improved color.

**Color printing processes.** All satisfactory color printing processes on paper or other reflecting support are subtractive. The pictures are formed by combining three positive dye or pigment images in yellow, magenta, and cyan. Chemical-toning, dye-mordanting, pigmented bichromated gelatin or silver halide-gelatin, and dye-bleaching processes have been used commercially. The imbibition process (Kodak Dye Transfer) is used for professional prints, three gelatin relief images (matrices) being made from the separation negatives, dyed in their respective color, and the dyes transferred in succession by contact to gelatin-coated paper.

Most color prints are made by coupler development using the monopack-type materials described for films, but on paper or white-pigmented cellulose acetate support. Color prints may be made by reversal processing to give prints from positive transparencies, or by the negative-positive process to give prints from color negatives, following essentially the techniques used with incorporated coupler films.

Dye-bleach processes have been proposed for many years, but in 1964 CIBA introduced Cilchrome commercially. In the dye-bleach processes dyes are incorporated in the emulsion layers and destroyed selectively by bleaching in presence of the negative silver image.

The Polaroid-Land Color Film, Polacolor, placed on the market in 1963, is a multilayer diffusion transfer system. The principle uses preformed dyes linked to developer molecules in the coatings. These combined molecules can wander by diffusion to a receiving sheet to form a picture, in the presence of alkali provided by a rupturable pod. Where development of the silver negative occurs, diffusion of the dyes is prevented, and a transferred positive color picture results.

**Color motion pictures.** Additive processes have been used for color motion pictures, one of which, the lenticular process, provided the first amateur 16-mm motion pictures.

Practically all color motion pictures are made by the subtractive processes, although chemical toning and dye toning were used earlier to some extent to give the colored images from color-separation negatives. The Technicolor process uses dye imbibition, the separation negatives being made in a beam-splitting camera. More modern processes use monopack films giving direct positives which may be printed onto a similar type of film to give duplicate positives, direct positives from which separation negatives are made for printing by dye imbibition, or in most cases, color negatives printed onto color print film, all using the incorporated coupler system. See COLOR; FILTER, COLOR.

[W. CLARK]

**Bibliography:** R. M. Evans, W. T. Hanson, Jr., and W. I. Brewer, *Principles of Color Photography*, 1953; J. S. Friedman, *History of Color Photography*, 1944; R. W. G. Hunt, *The Reproduction of Colour*, 1957; E. J. Wall, *The History of Three-color Photography*, 1925; see also PHOTOGRAPHY.

## Photoluminescence

A luminescence excited in a body by some form of electromagnetic radiation incident on the body. The term photoluminescence is generally limited to cases where the incident radiation is in the ultraviolet, visible, or infrared regions of the electromagnetic spectrum; luminescences excited by x-rays or  $\gamma$ -rays are generally characterized by special names. The graph of luminous efficiency per unit energy of the exciting light absorbed versus the frequency of the exciting light is called the excitation spectrum. The excitation spectrum is determined by the absorption spectrum of the luminescent body, which it often closely resembles, and by the efficiency with which the absorbed energy is transformed into luminescence.

Photoluminescence may be either a fluorescence or a phosphorescence, or both. Energy can be stored in certain phosphors by subjecting them to light or some other exciting agent, and can be released by subsequent illumination of the phosphor with light of certain wavelengths. This type of photoluminescence is called stimulated photoluminescence. In contrast to normal photoluminescence, which is constant in intensity as long as the intensity of the exciting light does not vary, stimulated photoluminescence decreases in intensity as the stored energy is released. See LUMINESCENCE.

[C. C. KLINK; J. H. SCHULMAN]

## Photolysis (photochemistry)

Chemical decomposition by the action of radiant electromagnetic energy, especially light. Photolysis occurs in certain crystals, notably the silver halides, when they are exposed to radiation. Here the effect of radiation is to produce a definite chemical change resulting in the separation of photolytic silver. Photolysis of the silver halides is discussed in this article because of the extensive investigations that have been carried out on these materials and because of their importance in the photographic process. Actually, photolysis occurs in many other materials, such as zinc oxide, the metallic azides and to a lesser extent in oxalates, styphnates, and fulminates.

A photographic emulsion consists of microcrystalline grains of silver bromide (AgBr) or silver chloride (AgCl) imbedded in gelatin. Upon prolonged exposure to light, so-called print-out specks of silver form within and on the surface of the grains. Shorter exposures produce a latent image which can be made visible by the process of development. Experiments have shown that the latent image is probably an early form of the print-out silver which results after long exposure.

**Gurney-Mott theory.** This theory of the photographic process proposes a two-stage mechanism as shown in Fig. 1. In the first stage a light quantum is absorbed at a point within the silver halide grain, releasing a mobile electron and a positive hole. These mobile defects diffuse to trapping sights (sensitivity centers) within the volume or

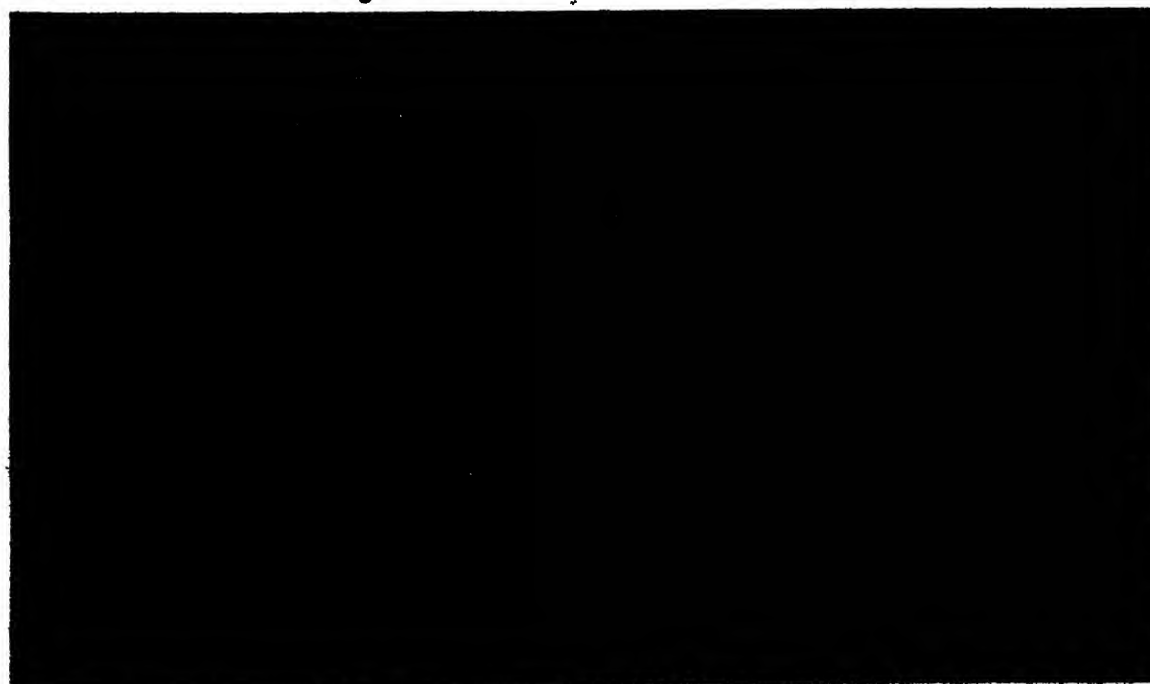


Unmasked negative



Part of corresponding negative carrying orange integral mask.

Positive made from the negative.



Color negatives and color positive of model of ethyl alcohol ( $C_2H_5OH$ ) molecule, showing relationship between the complementary colors. (Eastman Kodak)



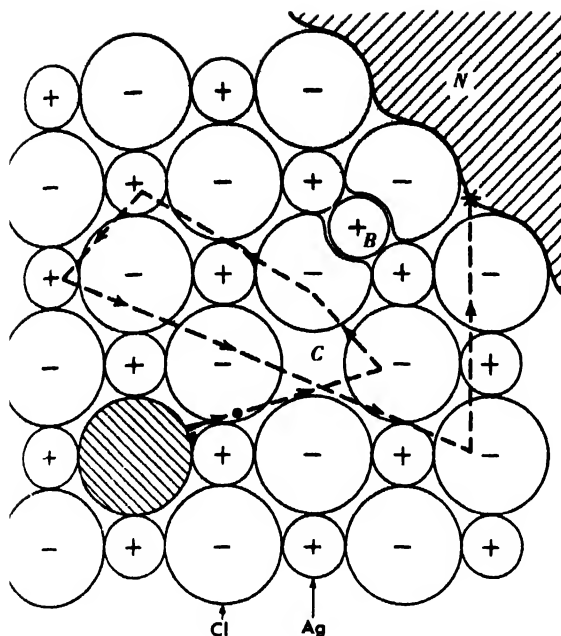


Fig. 1 Schematic representation of the Gurney-Mott theory. Dotted line shows the path of a photoelectron which is eventually trapped in the vicinity of a silver speck *N*. The interstitial silver ion *B* has come up to neutralize the charge of the electron and thus to add to the silver speck. (After J. R. Haynes and W. Shockley)



Fig. 2. Electronmicrograph of a silver halide emulsion grain exposed to repetitive pulses of light and to an electric field *E*. Cloud in the gelatin is thought to be due to escape of bromine from the grain, as if positive holes were displaced to the left by the electric field. (After J. F. Hamilton and L. E. Brady)

on the surface of the grain. In the second stage, the trapped (negatively charged) electron is neutralized by an interstitial (positively charged) silver ion, which combines with the electron to form a silver atom. The silver atom at the sensitivity center is capable of trapping a second electron, after which the process repeats itself, causing the silver speck to grow. It is assumed that the positive holes diffuse to the surface without recombining with electrons; at the surface they escape or react with the gelatin.

Criticism of the early Gurney-Mott theory as regards electron-hole recombination, the low density of interstitial ions, and the role of sensitizers has been given by N. F. Mott and J. W. Mitchell. However, the essential idea of an electronic process linking the initial absorption of light quanta with the formation of the image speck is well founded. Figure 2 shows an electronmicrograph of an emulsion grain after exposure to repetitive pulses of

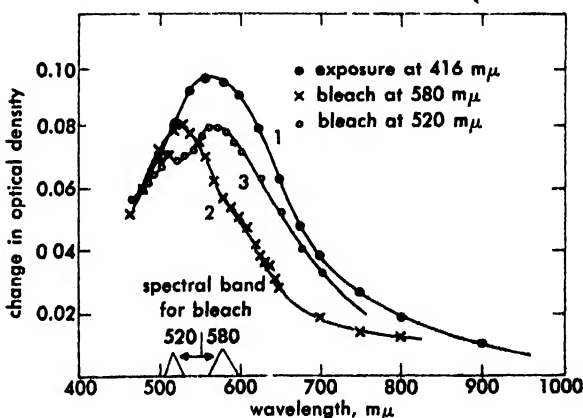


Fig. 3. Curve 1 shows the volume darkening produced at room temperature in a 0.37-cm thick AgCl crystal by absorption of  $4.2 \times 10^{16}$  photons/cm<sup>2</sup> at 416 mμ. Upon continued exposure, this darkening proceeds with high efficiency to a level which depends upon impurity content. Illumination within the band itself, however, produces bleaching, as shown by curves 2 and 3. Optical density is equal to  $\log_{10} I_0/I$ , where  $I_0$  is the intensity of incident light and  $I$  is the intensity of transmitted light. (After F. C. Brown and N. Wainfan)

light and to an electric field. The print-out silver specks occur on one side because of the action of the electric field, and there is experimental evidence for positive-hole migration in the opposite direction. Formation of the latent image in a fast emulsion containing sensitizers is highly complicated and is thought to involve the role of such structural imperfections as dislocations and jogs on dislocations. For additional information on latent image formation, see PHOTOGRAPHY.

**Large crystals.** The photolysis of large crystals of the silver halides has been studied in considerable detail. When interpreting the results on single crystals, it becomes necessary to distinguish between darkening produced by light in the volume of the crystal and darkening produced near the

surface. A pure crystal of AgCl will not darken appreciably upon exposure to light absorbed below the surface. On the other hand, crystals which contain small traces of impurity, particularly copper in the monovalent state, darken with high efficiency up to a saturation level which depends upon the amount of impurity present. Figure 3 shows the small amount of absorption for darkening of this type, which occurs in the early stages of exposure. Here the extinction of light arises primarily due to absorption. Prolonged exposure, however, produces darkening near the surface, which shows considerable light scattering. The centers responsible are similar to the colloidal metal particles formed in alkali halides by coagulation of *F*-centers during heat treatment. See COLOR CENTERS; PHOTOCHEMISTRY. [F.C.BR.]

**Bibliography:** S. Fujisawa (ed.), *Symposium on Photographic Sensitivity*, vol. 2, 1958; W. E. Garner (ed.), *Chemistry of the Solid State*, 1955; C. E. K. Mees, *The Theory of the Photographic Process*, 1954; J. W. Mitchell and N. F. Mott, The nature and formation of the photographic latent image, *Phil. Mag.*, [8] (21):1149-1170, 1957.

## Photometer

An instrument used in making measurements of light. In general, photometers may be divided into two classifications: (1) laboratory photometers, which are usually fixed in position, and (2) portable photometers, which are commonly used for photometric measurements outside a laboratory. Each of these two classes may be subdivided into visual photometers and photoelectric photometers. These in turn may be grouped according to function, such as photometers to measure luminous intensity, luminous flux, illumination, photometric brightness, light distribution, light reflectance and transmission, color, spectral distribution, and visibility.

**Visual photometers.** Most visual photometers use a Lummer-Brodhun photometer head or some adaptation of its principles. The Lummer-Brodhun photometer head is an optical device for seeing two sides of a white diffuse screen at the same time. On looking in the eyepiece the observer sees a circular or oval field composed of a pattern whereby adjacent parts are illuminated by light reflected from the two sides of the screen. Figure 1 illustrates the principle. Light from the two sources  $I_1$  and  $I_2$  to be compared falls on the white diffuse screen with sides  $s_1$  and  $s_2$  and is reflected in part to mirrors  $m_1$  and  $m_2$  and thence to the photometric prisms A and B. The central rays pass through the prisms while the outer rays are reflected by the prism. Concentric fields are thereby formed, an inner field by light from  $I_1$  and an outer field by light from  $I_2$ . The photometer is said to be balanced when both parts of the field are of equal brightness.

A more accurate type of Lummer-Brodhun photometer head is shown in Fig. 2. This differs only in the photometric prisms which are fashioned to

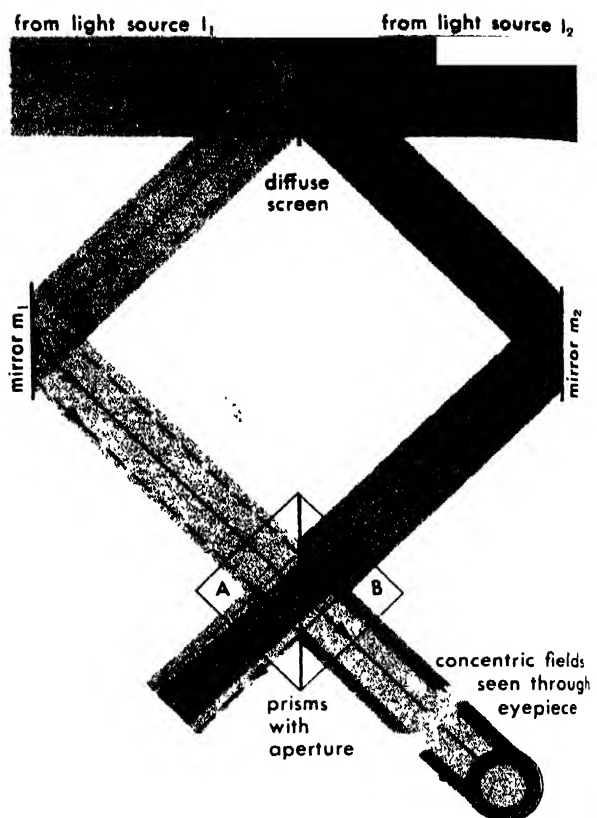


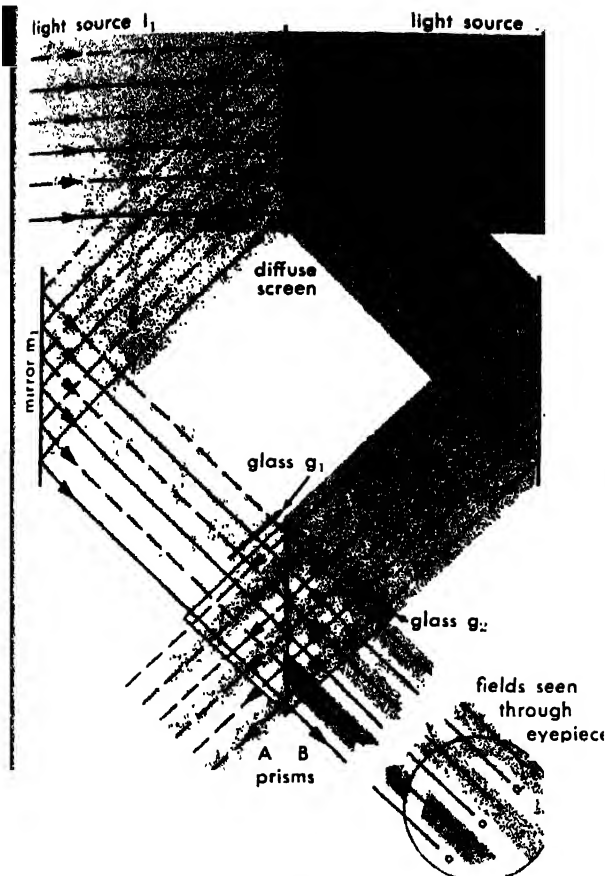
Fig. 1. Diagrammatic sketch of the Lummer-Brodhun equality-of-brightness photometer.

give a photometric field in which a and d are illuminated from light  $I_1$  and b and c from  $I_2$ . Clear glass strips  $g_1$  and  $g_2$  are added to absorb a slight amount of the light illuminating the fields b and d. This type of photometer head is balanced when parts a and c are equal in brightness (luminance) and when b and d are equal in contrast against their respective backgrounds. These Lummer-Brodhun photometer heads are incorporated in bar photometers, integrating spheres, distribution photometers, and portable photometers where visual photometry is used.

If there is a difference in color between the two light sources, difficulty is experienced in making a direct comparison, as in the Lummer-Brodhun principle, and a flicker photometer is often used.

**Flicker photometer.** In this photometer a single field is illuminated alternately by the sources to be compared. The position of balance is the point of minimum flicker. The rate of alternation is set fast enough that the color differences merge while the differences in luminance continue. Other systems of heterochromatic photometry are based on the cascade method, the compensation principle, and the use of filters.

**Illuminometer.** An illuminometer is a visual portable photometer. A typical illuminometer is shown in cross section in Fig. 3. Here the light entering is balanced against the light from the comparison lamp by means of a Lummer-Brodhun head



g. 2. Diagrammatic sketch of the Lummer-Brodhun contrast photometer.

is viewed through the eyepiece. The balance point is reached by moving the comparison lamp along the tube. A control box supplies a calibrated current to the comparison lamp and calibrated filters may be inserted in the light paths to extend the range of the instrument.

**Brightness meter.** This instrument for measuring photometric brightness, or luminance, may be either of the visual or photoelectric type. The visual type (Fig. 4) is a self-contained instrument operating on the same basic principle as the illuminometer except that in the brightness meter the balance point is reached by rotating a circular absorption gradient between the comparison lamp and the photometric head. The illuminometer may also be used to measure photometric brightnesses. The photoelectric type is also portable and consists essentially of a phototube, an amplifier to increase the current output, and microammeter calibrated to read directly the luminance of the surface viewed.

**Photoelectric photometers.** Photometers using barrier-layer cells and photoelectric tubes are classified as photoelectric photometers. Since the barrier-layer cells generate their own current and need no battery for operation, they can be used either in the laboratory or in portable instruments for use in the field. The phototube photometers, on the other hand, require a battery or other external power supply and are used mainly for laboratory

work. Both types of cells must be fitted with the correct filters so that they will duplicate the standard luminosity curve of the eye.

**Barrier-layer cell photometer.** This photometer is so portable and easy to use that it has almost monopolized field measurements since its introduction. It consists of a barrier-layer cell, which generates a small electric current (about 1 or 2 microamperes/foot-candle) when light or other radiant energy near the visible range falls on it, and a microammeter or galvanometer to measure this current. Modern barrier-layer cells generally use selenium as the light-sensitive material and are rugged in construction. The microammeter should have a low resistance (100 ohms or less), because the output of the cell is approximately linear only for low values of external circuit resistance. The photometer should be corrected to correspond to Lambert's cosine law of incidence (see PHOTOMETRY).

In this type of cell the current reaches its final value after a short time delay, due to the result of an adaptation effect, and also is influenced by temperature. Both effects will be a minimum with low external resistance. Figure 5 illustrates a portable type of barrier-layer cell photometer.

Barrier-layer cell photometers are extensively used as photographic aids to determine proper lighting and camera lens diaphragm openings (see PHOTOGRAPHY). For such applications the instrument is calibrated in an arbitrary scale. A simple manual computer is often included to permit converting the meter reading into the proper diaphragm opening and exposure time for a given type of film.

**Phototube photometer.** This photometer has as its elements a photoelectric tube, an amplifier to increase its sensitivity, a galvanometer or other suitable meter, batteries, and circuit arrangements to give the desired accuracy. In calibration the dark current, or the current flowing when the tube is not illuminated, is an important factor and should be subtracted or eliminated by circuitry.

**Integrating sphere.** Any of these photometers may be used with an integrating sphere to measure luminous flux. This is ordinarily made as an Ulbricht sphere (Fig. 6) although an octahedron or other shapes may be used. The inside surface has a white, diffusely reflecting finish which integrates the light. The light may come from a beam

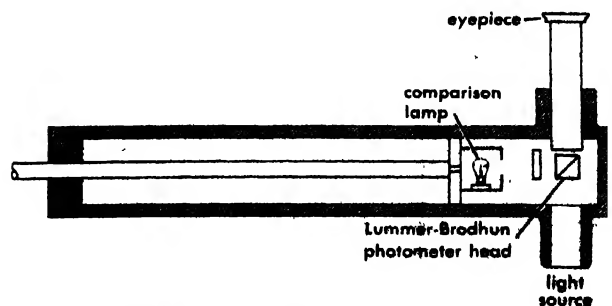


Fig. 3. Essentials of a Macbeth illuminometer.





Fig. 4. A self-contained visual brightness meter. (General Electric Co.)

projected into the sphere through an aperture or from a light source in the sphere itself. The integrated light is then measured by a suitable photometer head.

**Reflectometer.** This instrument, also called a transmissometer, combines integrating spheres and barrier-layer cells (Fig. 7). The reflectances of a surface may be determined by measuring the total light reflected from that surface when a beam of light strikes it, in comparison with the reflection from a standard surface. The transmittance can be measured by placing a sample of the material in the opening between the two spheres, one sphere containing the light source and the other the light-measuring cells.

**Distribution photometer.** The distribution photometer, or goniometer, measures the luminous in-

tensity at various angles from lamps, luminaires, floodlights, and searchlights. The light source can be moved to the desired angle and the light-measuring head (either visual or photoelectric) fixed, or the light may be fixed and the photometer made movable.

**Spectrophotometer.** Measurements of spectral energy from a light source are made by this device. It measures the energy in small wavelength bands by means of a scanning slip, and the results are presented as a spectral distribution curve. See SPECTROPHOTOMETRIC ANALYSIS.

**Visibility meters.** These operate on the principle of artificially reducing the visibility of objects to threshold values (borderline of seeing and not seeing) and measuring the amount of that reduction to determine the visibility by an appropriate scale.

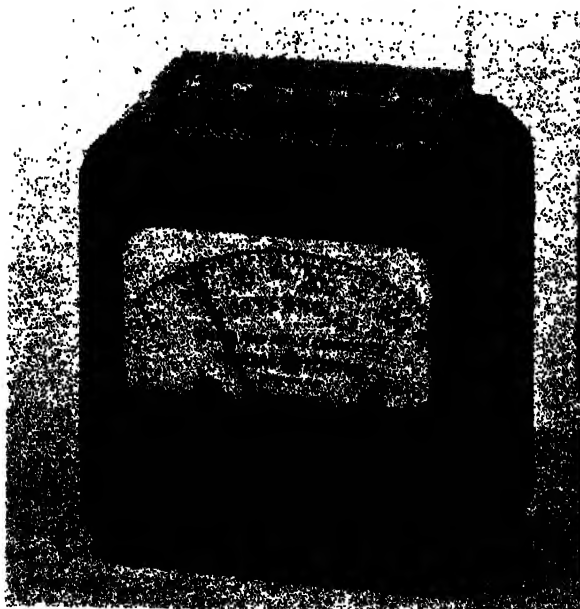


Fig. 5. A portable barrier-layer light meter with cosine and color correction. (General Electric Co.)

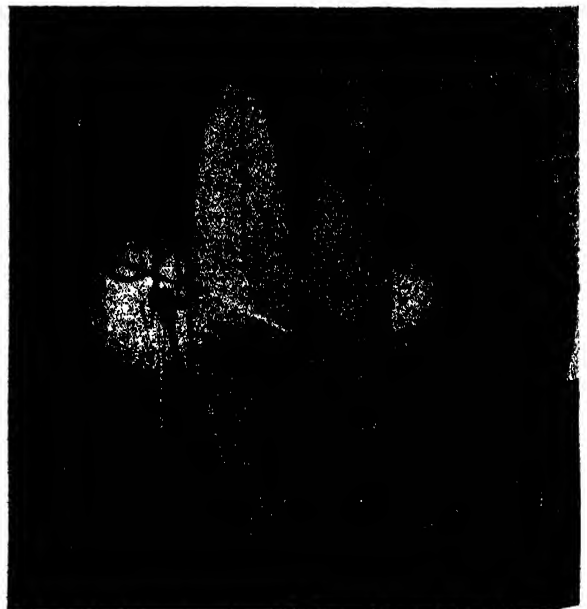


Fig. 6. Ulbricht sphere for measuring luminous flux and efficiency. (General Electric Co.)

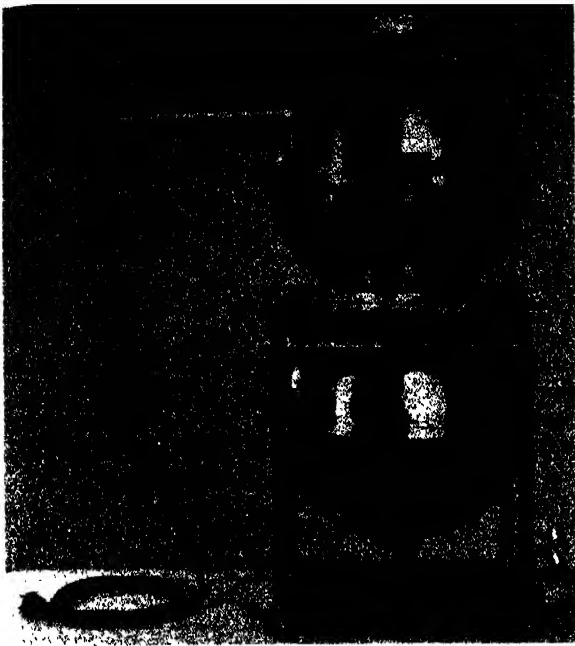


Fig. 7. A reflectometer-transmissometer showing the integrating spheres and barrier-layer cell photometer. (General Electric Co.)

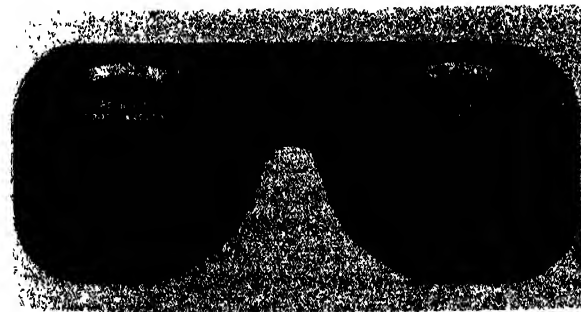


Fig. 8. A Luckiesh-Moss visibility meter. (General Electric Co.)

The Luckiesh-Moss visibility meter is probably the best known and easiest to use. This instrument (Fig. 8) consists of two variable density filters (one for each eye) that are adjusted so that a visual task seen through these filters becomes just barely discernible. Readings are on a scale of relative visibility related to a standard task.

The Cottrell visibility meter uses a luminous field superimposed by projection upon the visual target. The target is seen through a neutral-gradient filter which is adjusted until the target detail becomes barely perceptible, the results being expressed in terms of "contrast sensitivity."

Modern methods of automation have been applied to distribution photometers. Machines controlled by punched cards or tape give complete programming of the photometer, so no human attention during the test is necessary. The photometric data are automatically recorded and curves of luminous intensity can also be automatically drawn. [R.C.P.]

## Photometry

The branch of science that deals with the measurement of light. In general the instruments used in these measurements are called photometers, although other apparatus may be used in specialized instances, such as for the determination of wavelength.

Photometry is most usually concerned with measurements of luminous intensity (candlepower), luminous flux (such as light output), luminous flux density (illumination), luminance (photometric brightness), light distribution, color, spectral distribution, and the reflectance and transmittance of light. Photometry may even be extended to include visibility measurements. See ILLUMINANCE; LUMINANCE; LUMINOUS FLUX; LUMINOUS INTENSITY.

Since light is defined as radiant energy that is capable of producing visual sensation, photometric measurements either are made by the human eye or are based upon its visual responses. Because the spectral response of human eyes differs with individuals, the International Commission on Illumination (CIE) has adopted a standard luminosity curve, which has been accepted as being that of photopic vision of the normal eye. See VISION.

**Photometric laws.** Many photometric measurements are based upon the inverse-square law and Lambert's cosine laws of incidence and emission.

*Inverse-square law.* This law states that for a point source the illumination  $E$  on a surface varies directly with the luminous intensity  $I$  of the source and inversely as the square of the distance  $d$  between the source and the surface when the surface is normal to the light rays; that is,

$$\left(E = \frac{I}{d^2}\right)$$

In practical photometry where the source is a finite size, the error is less than 1% if the distance is more than 10 times the maximum dimension of the light source.

*Lambert's cosine law of incidence.* If the illuminated surface is not perpendicular to the light rays, the illumination on the surface varies as the cosine of the angle of incidence  $\theta$  between the normal to the surface and the incident ray. This law, together with the inverse-square law, gives

$$E = \frac{I}{d^2} \cos \theta$$

as the inverse-square law for any surface.

*Lambert's cosine law of emission.* This law is concerned with light sources and states that the luminous intensity in a given direction radiated or reflected by a perfectly diffusing plane surface varies as the cosine of the angle between that direction and the normal to the surface.

**Photometric measurements.** Photometry may be classified into two general classes, visual and physical. In visual photometry the eye is used directly in getting a photometric balance. In physi-

cal photometry a photoelectric cell, calibrated to response of the normal human eye, makes the measurement (see PHOTOMETER). In either case, the photometric measurements are eventually based on the primary standard of luminous intensity, which is called the new international candle in the United States and the candela in the rest of the world. See CANDLE.

This primary standard fulfills the requirements of being accurately reproducible with a radiation that follows definite laws as to both the amount of luminous energy emitted and its distribution throughout the visible spectrum. It is not, however, easy to use, so lamps, known as secondary standards, are carefully calibrated with the primary standard and used in its place. Other lamps, called working standards, are calibrated with the secondary standard, and are used in the laboratories for standards of luminous intensity and luminous flux.

The eye cannot measure the amount of intensity of light with any degree of accuracy because of its power of adaptation. It is a good judge of equality, however, when adjacent surfaces of much the same color quality are viewed simultaneously. In visual photometry, therefore, the eye views a photometric field that is made up of two parts, each of which is illuminated by one of the light sources to be compared. The photometer is then manipulated so that the two parts are equal in photometric brightness. The photometer is then said to be balanced. Balancing may be accomplished by (1) moving the photometer head in respect to the light sources, (2) moving one light source while holding the other source and the photometer head fixed, (3) introducing a disk of graduated density between one light source and the photometer head, or (4) otherwise adjusting the illumination on the two parts of the photometric field in a measurable way so that the photometer is balanced.

An example of the simplest type of photometric procedure is illustrated. A movable photometer head is mounted on rails between a working standard lamp and the test lamp, all three being in one straight line. This is called a photometer bench. The photometer head is moved until the luminances (photometric brightness) of the two sides of the head are the same. The luminance of a surface equals the illumination  $E$  times the reflectance  $\rho$ , or from the inverse-square law  $I\rho/d^2$ . Since the luminances are made equal

$$I_T \rho_T / d_T^2 = I_S \rho_S / d_S^2$$

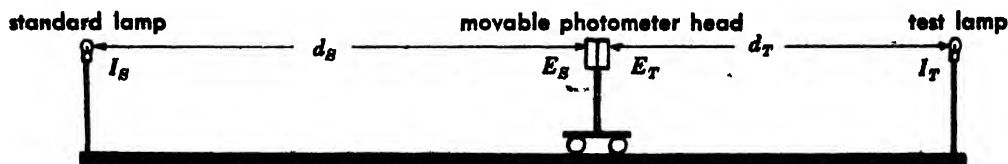
where subscript  $T$  identifies quantities relating to

the test lamp and subscript  $S$  identifies quantities relating to the standard. If the reflectances are equal, the luminous intensity  $I_T$  of the test lamp can be determined by

$$I_T = \frac{d_T^2}{d_S^2} I_S$$

Such visual methods are the basis of photometry and were the only ones used formerly, both in the laboratory and in the field. The accuracy of visual photometry is limited, however, because the response of one observer's eye may differ from that of another observer, as well as from the theoretical normal eye. The development of physical photometers in the last few years eliminates the errors caused by the uncertainty of the human eye. Physical photometers are quicker and easier to use, but they require frequent calibration, they vary with time, and they require special filters or other means of making their spectral response characteristics correspond to those of the theoretical normal eye of the standard observer. Both types of photometry are still used, but the physical photometer (especially using barrier-layer cells) is more popular and is used almost exclusively in field measurements of illumination.

**Photometric surveys.** Photometric tests are made of light sources, lighting units (called luminaires), lighting materials, and lighting installations. The measurements made of light sources and luminaires are of luminous intensity (directional candlepower, horizontal candlepower, spherical candlepower), luminous flux (output in lumens, efficiency), photometric brightness or luminance (foot-lamberts, candles per square inch), light distribution (distribution curves in one plane, iso-candle curves) and spectral distribution. The efficiency of a light source is generally expressed in lumens per watt. The efficiency of a luminaire is expressed as the ratio in per cent of the lumens output of a luminaire to lumens output of the light source within the luminaire. A photometric report on a luminaire can include candlepower distribution curves, zonal light flux distribution, total lumen output, efficiency of the luminaire, photometric brightnesses (luminances) at significant angles, coefficients of utilization and other data pertinent to the luminaire. The photometric tests of different types of luminaires (such as indoor lighting equipment, street lighting units, and searchlights) require individual photometric techniques, apparatus and methods of showing data, but they all follow the same principles.



Diagrammatic sketch of a photometric bench.

The photometry of a lighting installation includes the measurement of illumination (foot-candles), luminance, or the photometric brightness of surfaces, reflectances, and other quantities. This is called an illumination survey. The Illuminating Engineering Society has standardized various techniques of photometry, computation, and presentation of data for different classifications of lighting installations. The data often take the form of iso-foot-candle or isolux curves. See ILLUMINATION.

The measurement of visibility is connected with the photometry of a lighting installation, because it is one measure of the effectiveness of the lighting. Various types of visibility meters give readings of relative visibility based on threshold values (borderline of seeing and not seeing). See CANDLE-POWER; FOOT-CANDLE; FOOT-LAMBERT; LAMBERT; LUMEN; LUMEN-HOUR; LUMINOSITY FACTOR; LUMINOUS EFFICIENCY; LUMINOUS ENERGY; LUX; PHOT. [R.C.P.]

**Bibliography:** W. E. Barrows, *Light, Photometry, and Illuminating Engineering*, 3d ed., 1951; W. B. Boast, *Illuminating Engineering*, 2d ed., 1953; Illuminating Engineering Society, *General Guide to Photometry*, 1955; Illuminating Engineering Society, *I.E.S. Lighting Handbook*, 2d ed., 1952; J. W. T. Walsh, *Photometry*, 2d ed., 1953.

## Photon

An indivisible quantity of electromagnetic energy. J. C. Maxwell's theory of electromagnetic radiation successfully describes interference, diffraction, refraction, reflection at interfaces, and other phenomena associated with electromagnetic waves. The radiation has a dual character, however, and in other instances it displays aspects akin to those of particles, which are called photons. A photon is sometimes called a quantum or light quantum.

Photons are useful, for example, in understanding the Compton effect. Here electromagnetic radiation, usually in the form of x-rays, is scattered by electrons. The x-rays behave like photons of energy  $h\nu$  ( $h$  is Planck's constant;  $\nu$  is the frequency of the radiation) and of momentum  $h\nu/c$  ( $c$  is the velocity of light). In a photon-electron collision, momentum and energy are conserved, just as for particles in classical mechanics. The photon loses energy and has less momentum and therefore a longer wavelength after the collision. The electron gains the energy and momentum that the photon loses.

The photon aspect of radiation is also clearly evident in the phenomenon of photoemission. In the simplest case, a single electron absorbs the entire energy of an incident photon and appears outside the emitter with a clearly defined maximum energy described by the Einstein photoelectric law. A photon can carry away angular momentum when it is emitted during an electronic transition in an atom. See COMPTON EFFECT; ELEMENTARY PARTICLE; PHOTOEMISSION; QUANTUM (PHYSICS); QUANTUM MECHANICS. [L.A.]

## Photoperiodism in plants

A variety of growth and development responses of plants to daily duration of light and darkness. Although photoperiodism occurs in many species from the lowest to the highest, it was first observed in angiosperms, and the study of this group has provided most of the present knowledge (see ANGIOSPERMAE). In 1920, W. W. Garner and H. A. Allard discovered photoperiodism in plants. They reported that plants, such as Biloxi soybean and Maryland Mammoth tobacco require short days and long nights for flowering, that other plants need long days and short nights, whereas still other plants were seemingly uninfluenced by day length. Subsequently, they found that flowering was not the only day-length-regulated response of plants, for example, several phenomena such as production of bulbs, tubers, and runners, coloration of some kinds of fruits and leaves, formation of buds, and the development of bud dormancy are also controlled by day length.

The phenomenon was called photoperiodism because it was concerned with effects of light on the plant in relation to time. The time measure was a daily one, being the period either of light or of darkness, or a ratio of the one to the other. In a given latitude the daily periods of light and dark vary with the season, and most plants respond accordingly. Timing of seasonal responses by duration of light is much more constant from year to year than timing controlled by temperature, rainfall, or some randomly variable condition.

**Terminology.** Certain terms are commonly used in discussing photoperiodism of flowering plants.

**Short-day plant.** This kind of plant flowers best, or only, when daily light periods are short and dark periods long. Examples are poinsettia, soybean, cocklebur, and chrysanthemum.

**Long-day plant.** This kind of plant flowers best, or only, when daily light periods are long and dark periods are either short or absent. Examples are wheat, oats, barley, beet, and poppy.

**Indeterminate, or day-neutral, plant.** The flowering of this kind of plant is independent of daily duration of light and darkness. Examples are tomato and garden bean.

**Photoperiod-induction treatment.** The day-length, exposure to light, which is followed by the appearance of microscopic flower buds (flower primordia).

**Critical day length or night length.** This is the day or night length which differentiates between inductive and noninductive photoperiods for either long- or short-day plants.

When photoperiodism was first discovered, the specific roles of light and dark in the process were unknown, but it has since been shown that the time-measuring processes occur during the dark periods. Short-day plants are thus more correctly long-night plants, and long-day plants, short-night plants. Because of long usage, however, the terms short-day and long-day will probably be retained.

**General description.** Information descriptive of photoperiodism is available for a wide range of plants and from it can be gathered certain general facts. For example, critical day lengths vary with species or even varieties, and day lengths favorable to flowering of long- and short-day plants frequently overlap. Thus, a short-day cocklebur may flower on photoperiods of  $15\frac{1}{2}$  hours or less and long-day *Hyoscyamus* on photoperiods of 12 hours or more. Plants also differ markedly in the number of flower-promotive photoperiods they require for induction. For example, cocklebur needs only one such photoperiod under ideal conditions, Biloxi soybean needs two or three, and chrysanthemum requires several.

The leaf blade is the portion of the plant that reacts effectively to photoperiods that promote flowering. A single leaf exposed to long dark periods may induce flower bud formation on a short-day plant even though the rest of the leaves of the plant are on short dark periods. This finding shows that flowering results from the action of a positive flower-inducing stimulus generated in the leaf during long dark periods rather than from the suppression of a stimulus for vegetative development by long dark periods.

Although flowers form at a distance from the leaves, the flowers develop in response to something that happens in the leaves during the dark period. A flower-promoting product generated in the leaf, possibly of hormonal nature, apparently moves to the growing points where it controls differentiation of the flowers (see PLANT GROWTH; PLANT HORMONES). Migration of enough of this product from the leaf to the growing point to effect floral induction in many plants requires that the leaf remain attached to the plant for several hours of the photoperiod following the long dark period. In cocklebur, for example, no flowers form if the exposed leaf is removed at the close of the long dark period, but flowers form abundantly if the leaf is not removed until about 8 hours later.

Darkness must be complete to be effective on photoperiodically responsive plants; a very small amount of light nullifies its action. Illumination greater than 0.02 ft-candle of incandescent-filament light throughout the dark period inhibits flowering of several kinds of short-day plants. The dark period, moreover, must be continuous to be effective; an interruption near its middle as brief as 1 minute and with light energy equivalent to less than 50 ft-candle minutes of incandescent-filament light prevents flowering of plants such as poinsettia and soybean. The brief period of light thus has consequences that continue for several hours.

**Nature of the light reaction.** The various responses of plants to day-length treatment are descriptive of photoperiodism, but knowledge about them does not immediately lead to an explanation of the processes. Although the biochemical reactions involved in flowering are complex and are almost completely unknown, it is probable that the chains of reactions leading to flowering of a pho-

toperiodically sensitive plant and an indeterminate one are very similar. The chief difference lies in the fact that the course of the reactions, as judged by the flowering response, can be completely reversed in the photoperiodically sensitive plant but not in the indeterminate plant by a light reaction of very low energy.

The nature of this light reaction has been studied in considerable detail, and some progress has been made in understanding how it works. The most significant and, to the experimenter, the most useful information about the reaction is that it is reversible by radiant energy. It goes in one direction if the leaf is irradiated with red light (about 6500 Å) and in the other if the leaf is irradiated with so-called far red, or near-infrared, light (about 7350 Å). If given in the middle of a long night, red light acts to prevent flowering of short-day plant and to promote flowering of long-day ones. Far red in each case counteracts the action of red.

**Action spectrum of photoperiodism.** The maximum effectiveness of each of these reactions is at the wavelengths mentioned in the preceding section, but other wavelength regions of the spectrum are also active. For example, the relative effectiveness of various wavelengths of light in causing the reaction that has its maximum near 6500 Å is

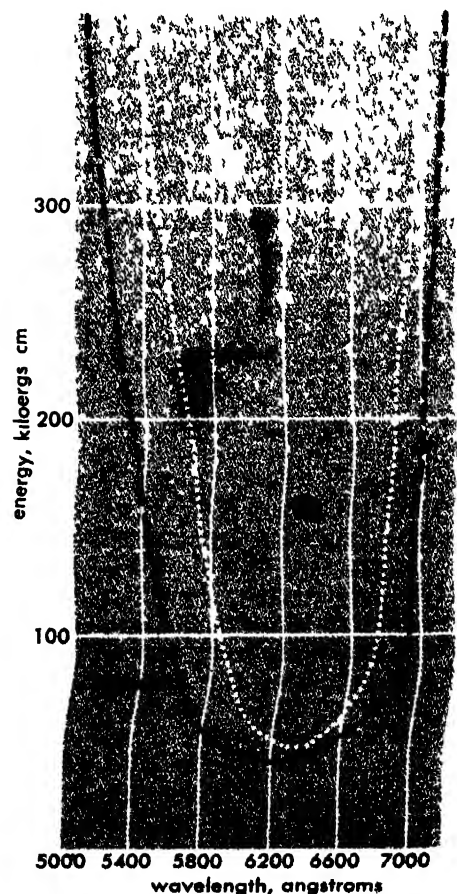


Fig. 1. Action-spectrum curves showing energy required for inhibition of flowering of two short-day plants, soybean and cocklebur.

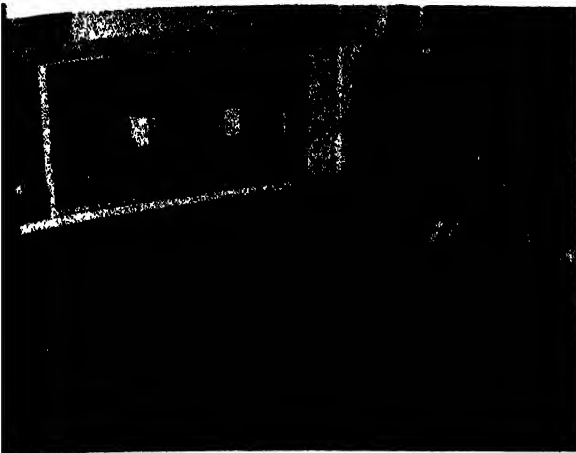


Fig. 2. View of spectrograph with front removed to show part of optical system.

shown in Fig. 1. Such curves, which relate the energy of incident light for a given response to wavelength, are called action-spectrum curves. The data for the curves come from carefully performed physiological experiments in which light of known wavelength and energy is applied to plants in the middle of the long nights to prevent or promote flowering according to the photoperiodic response of the plant. Several energy levels, some too high and some too low, are given to different lots of plants at each wavelength region so that a reliable estimate can be made of the minimum energy required for response at each region.

Light of high spectral purity is produced by several different methods. It may be obtained by passing the light from a carbon arc through a two-prism spectrograph (Fig. 2), which, at a distance of about 40 ft, produces a spectrum 3 in. or more high and about 6 ft in width from the violet to the red. Leaves or whole plants are given their dark-period interruptions by placing them along this spectrum at fixed stations for which the energies and wavelength ranges are known (Fig. 3).

Action-spectrum curves sometimes give a clue to the nature of the compound that is responsive to light in the photochemical reaction. Special effectiveness of light from certain parts of the spectrum, such as that of the red for the photoperiodic reaction, indicates that a pigment is present that actively absorbs red light. Under ideal circumstances, the action spectrum for the response may closely resemble the absorption spectrum of the active pigment as to give information about the biochemical nature of the pigment. Because the effective region of the spectrum for the photoperiodic reaction is mainly in the red, it can be concluded that the active pigment is probably blue. However, blue-pigmented compounds having absorption spectra closely similar to the photoperiodic action spectra have been obtained from photoperiodically responsive plants. Phycocyanin, a straight-chain tetrapyrrole which occurs abundantly in blue-green algae, has an absorption spectrum of the type indicated by photoperiodic action spectra (see CYA-

NOPHYTA). Blue-green algae, however, are not known to be sensitive to photoperiod, and photoperiodically sensitive higher plants are not known to contain phycocyanin. A compound of similar nature, nevertheless, is the most probable form of the pigment.

The several action-spectrum curves of long- and short-day plants are essentially alike. This similarity indicates that the primary photoreaction in the two kinds of plants is the same even though the effects of identical treatments are diametrically opposite. Intergrafting of long- and short-day plants before identity of their action spectra was established indicates that either type, held on day lengths promotive of its own flowering, could induce flowering in a graft partner of the opposite day-length type even though the latter was subjected to photoperiods not promotive of flowering. Such experiments show identity in details of the flowering process in long- and short-day plants; the action-spectrum experiments show that the identity extends to the preceding light reaction that regulates these details.

**Photoresponses.** Comparison of action spectra is a reliable method of establishing the identity of reactions that cause different plant responses, an important function of these action-spectrum measurements. This method is used to show that the photoreaction that regulates flowering of long- and short-day plants also regulates several other plant responses. Although the latter are not commonly regarded as photoperiodic, they are mentioned briefly here because of the contribution that knowledge about them makes to an understanding of photoperiodism. Among these responses are the germination of light-sensitive seeds, some features of the coloration of fruits upon ripening, and variations in the amount of growth in length made by the internodes of some kinds of plants.

The germination of seeds, such as those of Grand Rapids lettuce, is promoted by light, especially red light. The action spectrum for promotion of germination is so similar to the action spectra for regulation of flowering that the light reaction can be

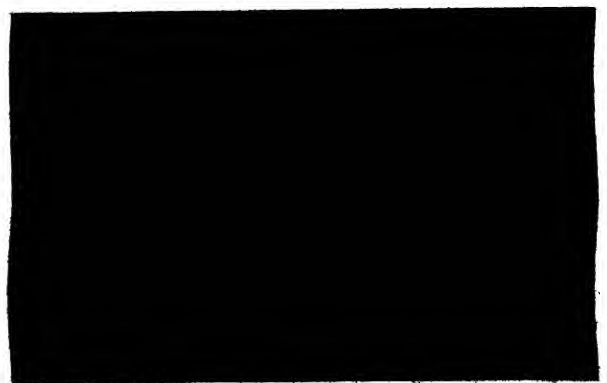


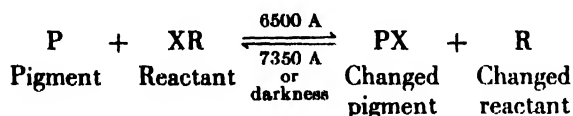
Fig. 3. Leaves of soybean plants in focal plane of spectrograph. All leaves except terminal leaflet of third compound leaf were removed prior to beginning of experiment.



assumed to control both. That this is true is shown by the occurrence of reversibility of the germination response at the same wavelength regions that cause reversal of flowering; in Grand Rapids lettuce far-red light reinhibits the germination of seeds which are induced to germinate by red light.

A single brief light treatment suffices to induce the germination of lettuce seeds; repeated treatments at daily intervals, as practiced in the control of flowering of many plants, are not required. For some seeds, however, light treatments for several hours or repeated daily treatments are required, and certain workers have proposed that these seeds exhibit true photoperiodic response. This view is probably correct; at least, the presence of the reversible photoreaction in them has been demonstrated, but complete details of the photoperiodic response in seeds remain to be elucidated.

**The photoreversible reaction.** Knowledge about the reversible photoreaction, drawn not only from photoperiodism but also from studies of other phenomena in which the reaction occurs, leads to the following general expression:



The pigment form (PX) resulting from the action of red light probably regulates the various biological responses in which this reaction occurs. Evidence for this comes from the germination of seeds that have long been in darkness in the imbibed but dormant condition; a very low energy of red light, which would change a small amount of P to PX, induces their germination.

Inclusion of XR and R in the reaction is forced by the fact that the relative sensitivity of the system to red and far-red-radiant energies may be widely different in different species or even in the same species under different experimental conditions. These changes in sensitivity are usually of a reciprocal nature in which an increased sensitivity to red is accompanied by an opposite change in sensitivity to far red. Such variations would result from changes in amount of reactant present. Change in sensitivity of the reaction to red and far red has been examined in most detail in seed-germination studies but has been observed to occur also in photoperiodism.

In both seed germination and flowering there is evidence that the active pigment may undergo change of form in darkness. In both instances, this change during darkness appears to be from the far-red-absorbing (PX) to the red-absorbing (P) form. In lettuce seeds more than 24 hours appear to be required for effective change, whereas in flowering the period seems to be much shorter. Reasons for the difference are possibly connected with differences in the fractional amount of the total pigment in the active form required to regulate the two processes.

Elongation of internodes of plants, another photoreversible phenomenon, depends markedly upon the kind of light to which the plants are exposed at the beginning of the daily dark period. Internodes of pinto bean plants that receive far red at this time often elongate to more than three times the length of those that receive red. This reaction, as in seed germination and flowering, is repeatedly reversible with red and far red. The three phenomena are thus controlled by the same photochemical reaction. The response of internodes to red-far-red irradiation treatments shows that the pigment system is present and functional in the plant immediately after the light is discontinued, a point not readily shown by the photoperiodic reaction in its regulation of flowering.

Most beans are said to be day-neutral with respect to flowering. Therefore, it is of special significance that the red-far-red reaction is shown to operate in the bean plant to control response other than flowering. Similar findings have been made with tomato, which also is day-neutral in photoperiodic response. The photoreversible red-far-red reaction has nevertheless been shown to regulate seed germination, internode growth, and fruit coloration in tomato. The reaction is thus clearly shown to be present and operative in plants such as bean and tomato; the reason for its failure to regulate flowering in these plants is not yet understood.

The fact that the processes leading to flowering are under control of a reaction that can be started and stopped at will opens the way to studies of the rates of those reactions. Thus, one can irradiate short-day plants in the middle of dark periods that would otherwise be promotive of flowering and thereby start reactions that lead to its inhibition. The speed with which such reactions proceed is tested by giving far red at various intervals after the red. If given immediately, far red almost completely counteracts the inhibitory effect of red, but if given 30–60 minutes after red, it is ineffective, failing completely to reinduce flowering. This is taken to mean that products of reactions favorable to flowering that accumulated during the first half of the dark period are rather completely destroyed by irradiating the plant with red and leaving it in that condition for approximately 30 minutes.

Information has thus been obtained about the photoreaction of photoperiodism by physiologic experiments even though the beginning and end compounds of the reaction are still unknown. The active pigment is not chlorophyll, and the concentration of the active pigment is known to be extremely low because it imparts no visible color to certain albino plants which nevertheless exhibit photoreversible response and therefore contain the pigment. Therefore, the reaction is one in which profound growth effects result from exceedingly small amounts of photoreactive pigment irradiated by very low light energies. The amount of pigment actively changed by the light into the form that regulates growth is exceedingly small, indicating



that it probably operates as an enzyme or other energy-transferring device.

The active pigment, phytochrome, was detected in live tissue by a special spectrophotometric procedure in 1959 and was extracted but not yet purified and identified by early 1960. Further study of the extracted pigment may clarify the nature of the primary photoreaction and lead to an understanding of the biochemical steps involved in flowering and other photoperiodic responses. The latter does not necessarily follow, however. Identification of chlorophyll as the photoactive pigment in photosynthesis, for example, did not markedly improve understanding of photosynthesis. The extraction of phytochrome, however, reopens the subject of photoperiodism and related photoresponses to studies of a type not previously possible.

Many workers have shown effects on flowering of auxins, antiauxins, and various other kinds of compounds that influence plant growth and flowering. Very little is known about the way these substances operate to cause their effects. Indeed, whether the effects result from the direct participation of these substances in the flower-regulating reactions or from their more general influence on growth is not clear. Nevertheless, various workers are currently doing pertinent experiments involving applications of the chemical treatment and the photoinductive dark period in variously timed sequences. Such experiments will at least indicate the order in which the chemical treatments and the photoregulated reactions occur in causing their combined effects and may reveal further facts about either the reactions leading to flowering or the ways in which these various growth-regulating substances operate in the plant. See PLANT HORMONES.

**Applications of photoperiodism.** Although the mechanism of photoperiodism is far from completely understood, much practical application is made of the descriptive knowledge about the phenomenon. Plant scientists working in plant breeding, plant pathology, or other fields use artificially controlled photoperiods to regulate the time of blooming of their plants. The knowledge is also used by growers of certain crop plants, particularly ornamental ones, to bring their plants into bloom out of season so as to take advantage of favorable market conditions. The production of chrysanthemums in the United States depends largely on artificial control of light and darkness, a practice that permits their production throughout the year instead of during a few months in autumn (Fig. 4). Other ornamentals in which the use of controlled day length plays a practical part are asters, tuberous-rooted begonias, and orchids.

Distribution of plants in nature depends to a considerable extent on their photoperiodic responses. Therefore, knowledge about photoperiodism has served to clarify certain ecological problems. The onset of bud formation in a great many woody plants is determined by the seasonal change in daylength, and in some species occurs rather early in the summer. Other environmental variables proba-

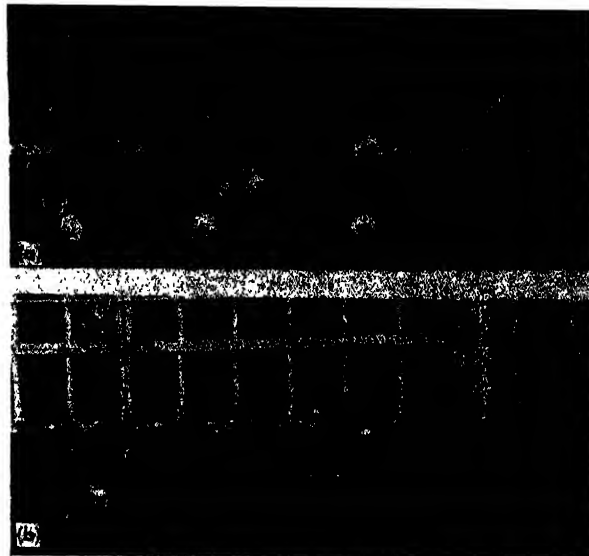


Fig. 4. (a) Plants of the common China aster, *Callistephus chinensis*, grown in a northern greenhouse in the short days of midwinter. (b) Plants of the same kind and age grown under the same conditions, except that their daylight period was lengthened 4 hours by artificial lighting.

bly play parts in causing the onset of dormancy in some plants, but in many woody plants, growth stops and buds form long before the arrival of low night temperatures to which these events are sometimes erroneously attributed.

Shoot elongation, which is stopped by short photoperiods in many species, is not resumed in some kinds when they are returned to long days. Thus, plants of *Catalpa bignonioides* stop producing new leaves as soon as they are subjected to short photoperiods and, if held on short days for a month or more, they do not resume growth when returned to long days unless the temperature is sharply lowered for several days. *Weigela* shrubs, on the other hand, resume growth immediately without any low-temperature treatment when they are restored to long days. The inhibitory effects on growth apparently occur in the leaves because removal of the leaves of *Weigela* is immediately followed by resumption of bud growth even though the plant continues on short photoperiods. The new growth, however, is quickly stopped if the short days are continued. See PLANT PHYSIOLOGY. [H.A.BO.]

**Bibliography:** A. Hollaender (ed.), *Radiation Biology*, vol. 3, 1955; A. E. Murneek et al., *Vernalization and Photoperiodism*, A symposium, *Lotsya, A Biological Miscellany*, vol. 1, 1948; D. Rudnick (ed.), *Aspects of Synthesis and Order in Growth*, 1955.

### Photophore gland

A highly modified integumentary gland which arises from an epithelial invagination into the dermis. It becomes cut off from its site of origin and develops into a luminous organ comprised of a lens and a light-emitting gland, back of which is a vis-

mented reflector of probable dermal-cell origin. These luminous bodies occur in deep-sea teleosts and elasmobranchs, which live in areas of total darkness. See CHROMATOPHORE; EPITHELIUM; GLAND. [O.E.N.]

## Photoreception

The process of absorption of light energy by plants and animals, and its utilization for biologically important purposes. In plants, photoreception plays an essential role in photosynthesis and an important role in orientation. Photoreception in animals is the initial process in vision. See PHOTOSYNTHESIS; TAXIS.

The photoreceptors of animals are highly specialized cells which are light-sensitive because they contain pigments which are unstable in the presence of light. These light-sensitive receptor pigments absorb quanta of radiant energy and subsequently undergo physicochemical changes, which are translated into nerve impulses conducted to the central nervous system.

**Morphology of photoreceptors.** While their gross structures differ widely, photoreceptors examined with the high-resolution electron microscope appear to have in common a fine structure featuring the presence of membranous organelles with appreciable surface area.

**The vertebrate eye.** The vertebrate eye (Fig. 1) can be exemplified by the human eye. The distal segment of the vertebrate retinal rod cell (Fig. 1d) consists of a stack of disks enclosed by a membrane. In the rods of the guinea pig, rabbit, and mouse these disks consist of two membranes 30 Å

thick. These enclose a space of about 80 Å, and the disks are spaced 100–200 Å apart. Each distal segment contains many disks; in the frog, the estimated number is 1000. The disks in the distal segment are attached to a cilium which runs the length of the segment and connects the distal and proximal segments of the rod cell (Fig. 1d). In the cone distal segment of the perch, the disks are composed of a single membrane 170 Å thick. Minor variations in dimensions of the fine structure occur in the different species of vertebrates examined.

**Arthropods.** Photoreceptors of arthropods are superficially quite different from vertebrate eyes; but the fine structure of receptor cells reveals some similarities (Fig. 2a,b). Of particular interest are those portions of the retinula cells, the rhabdomeres, which are membranous organelles of appreciable surface area. The rhabdomeres of the *Limulus* eye are composed of interdigitating microvilli from the apposed surfaces of the membranes of the retinula cells in the central region of the ommatidium (Fig. 2c,d). Collectively, the rhabdomeres constitute the rhabdome, in the center of which is the dendrite of the eccentric sense cell. The rhabdome is situated beneath the crystalline cone directly in the light path. The rhabdome of the eyes of spiders and centipedes is similar to that of *Limulus*, with minor variations. In the eye of the fly, the rhabdomeres are located along the central margins of the retinula cells and constitute the rhabdome, which is also centrally located and readily accessible to any light entering the lens. These rhabdomeres consist of stacks of hexagonal tubes, the long axes of which are parallel to

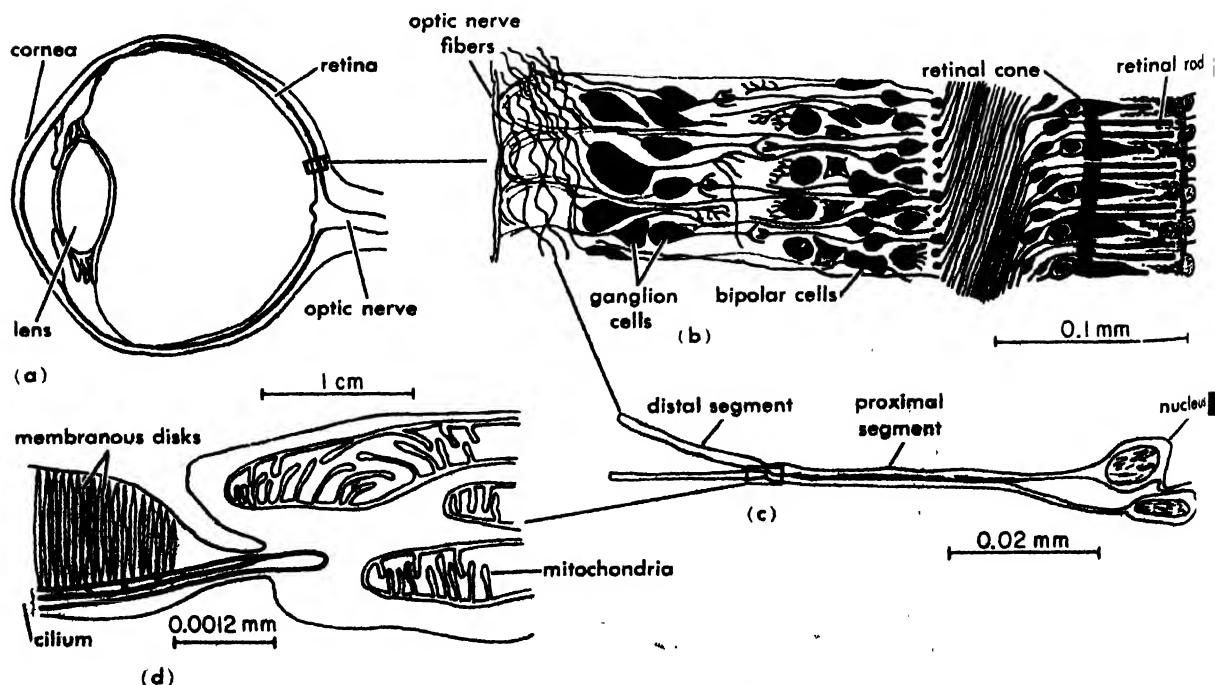


Fig. 1. Gross and fine structure of photoreceptors of a typical vertebrate. (a) Section through eye. (b) Layers of retina (from S. L. Polyak, *The Retina*, Univ. of Chicago Press, 1941). (c) Retinal rods (from S. L.

Polyak, *The Retina*, Univ. of Chicago Press, 1941). (d) Rod (from F. DeRobertis, *J. Biophys. Biochem. Cytol.*, 2(3):319–329, 1956).

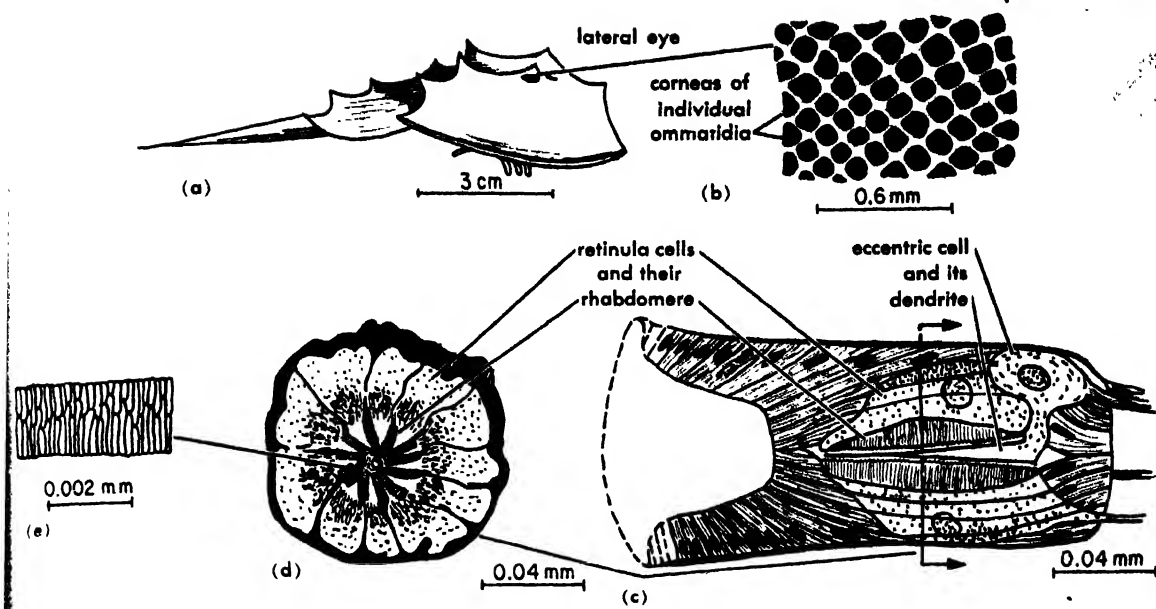


Fig. 2. Gross and fine structure of photoreceptors of the horseshoe crab, *Limulus*. (a) *Limulus*. (b) Eye. (c) Longitudinal section through ommatidium. (d) Cross section through ommatidium (from W. H. Miller, Ann.

N.Y. Acad. Sci., 74(2):204-209, 1958). (e) Microvilli of rhabdomere (from W. H. Miller, Ann. N.Y. Acad. Sci., 74(2):204-209, 1958).

radius which bisects the retinula cell. The hexagonal tubes are 370 Å in diameter and consist of a peripheral membrane about 120 Å thick. Some variations in basic structure exist in other arthropods.

**Mollusca.** Molluscan photoreceptors also possess membranous organelles containing appreciable surface areas. The eye of the squid contains fused rhabdomeres which appear to be composed of tubules. That of *Pecten* (Fig. 3) has a unique membranous organelle consisting of concentrically arranged light and dark bands which together make up an oval or spherical body of about 1 μ diameter. The dark bands of this organelle are about 50 Å thick, and appear to be continuous with the stalks of cilia which are attached to basal bodies within the cytoplasm of the receptor cell.

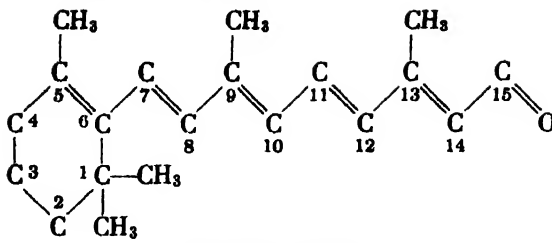
**Common characteristics.** The photoreceptors described have in common their accessibility to light, the extensively developed and intricate membranous organelles and some other features. Extensive membranous organelles have also been observed in the photoreceptors of flatworms *Planaria*, in the eye spot of the protozoan, *Euglena*, and in the chloroplasts of green plants. It has been postulated that these organelles of photoreceptors contain light-sensitive pigments which are thus effectively exposed to any incident radiant energy. Unfortunately, direct proof of this postulate is lacking. The visual pigment of vertebrate rods, called rhodopsin, has been localized in the distal segment of the rods; but there is no evidence as to where in the distal segment the pigment is situated.

**Chemical behavior of photosensitive pigments.** The light-sensitive pigments of photoreceptors, which are responsible for the absorption of radiant

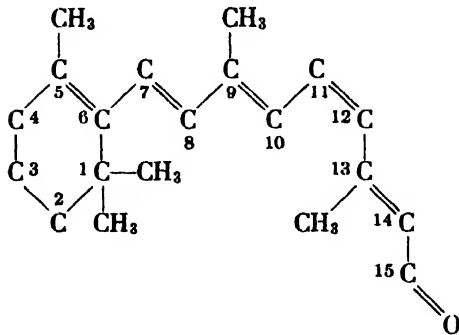
energy, all appear to have a similar chemical constitution. Rhodopsin, the pigment of vertebrate retinal rods, consists of a protein conjugated with a carotenoid. The latter has been identified as retinene,  $C_{20}H_{28}O$ , the aldehyde of vitamin A. The absorption maximum of free retinene is at 380 mμ; when it is combined with the protein, opsin (molecular weight about 40,000), the visual pigment rhodopsin is formed, and the absorption maximum shifts to about 500 mμ. The agreement of the absorption spectrum of rhodopsin with the spectral sensitivity of human dim-light vision and with that of many other vertebrates clearly implicates rhodopsin in light reception by the rods.

**Ultrastructure and rhodopsin.** Some evidence suggests that rhodopsin forms an integral and highly organized part of the ultrastructure of the rod. The pigment contributes 40% of the weight of frog-rod outer segments. Polarized light, fluorescence, and density studies all suggest that the molecules are highly orientated within the rod. Also, the rhodopsin molecule has a phospholipid component which, if totally removed, destroys the stability of the pigment. The interesting possibility exists that this phospholipid fraction may represent an actual part of the structural fabric of the rod to which the rhodopsin molecule is bound in the distal segment.

**Retinene isomers.** The bleaching and resynthesis of rhodopsin depend upon the existence of several retinene isomers. The most important pair are illustrated in the structural formula. Only 11-*cis*-retinene will, when incubated with opsin, form rhodopsin. Illumination of a rhodopsin solution results in the release of retinene from the protein in the all-*trans* form. The mechanism of bleaching



All-trans retinene



11-cis-retinene

Structure of two retinene isomers

thus involves (1) the absorption of a quantum of light energy by the molecule, (2) the isomerization of the 11-*cis*-retinene to all-*trans*-retinene, and (3) the thermal hydrolysis of this incompatible isomer from its site on the protein. The intermediate in this process, namely, the complex between

opsin and all-*trans*-retinene before the latter has hydrolyzed away, exhibits an absorption spectrum displaced slightly towards shorter wavelengths from that of rhodopsin, and has been called metarhodopsin.

**Dark adaptation.** After exposure to light, a recovery of sensitivity by the receptor cell takes place; this process is referred to as dark adaptation. It is dependent upon (1) an adequate supply of 11-*cis*-retinene and (2) the rate of recombination of opsin with this active retinene isomer. 11-*cis*-Retinene apparently is supplied partially by the action of an enzyme, retinene isomerase, upon the all-*trans*-retinene released from bleached rhodopsin. In addition, an equilibrium exists [catalyzed by alcohol dehydrogenase and diphosphopyridine nucleotide (DPN)] between retinal vitamin A and retinene. Opsin traps the 11-*cis*-retinene which is formed to reconstitute the visual pigment.

**Other pigments.** In addition to rhodopsin, several other visual pigments have been identified in the retinas of vertebrates. Fresh-water fishes, some amphibians, and certain reptiles possess a slightly different form of retinene, called retinene<sub>2</sub>. This, when combined with rod opsin, forms porphyropsin, a visual pigment similar to the rhodopsins, but with an absorption maximum at 522 mμ. The protein moiety of visual pigments, too, can be altered. Cones contain a different opsin which combines with retinene<sub>1</sub> to form iodopsin. This pigment, which has been isolated from the chicken retina.

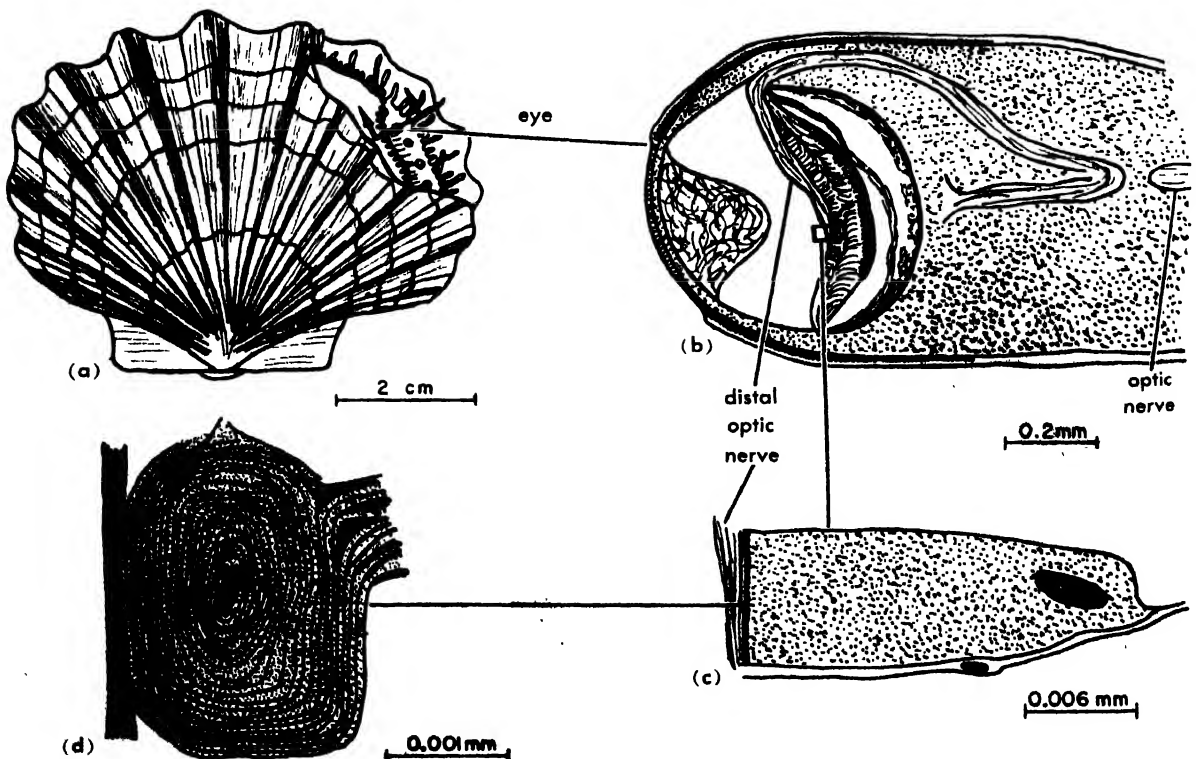


Fig. 3. Gross and fine structure of photoreceptors of the scallop, *Pecten*. (a) *Pecten*. (b) Longitudinal section through eye (from W. J. Dakin, *Quart. J. Microscop. Sci.*, 4:49-112, 1910-1911). (c) Distal sense cell (from

W. J. Dakin, *Quart. J. Microscop. Sci.*, 4:49-112, 1910-1911). (d) Appendage (from W. H. Miller, *Ann. N.Y. Acad. Sci.*, 74(2):204-209, 1958).

differs from rhodopsin in its absorption maximum (562  $m\mu$ ) and also in its much faster regeneration rate from cone opsin and 11-*cis*-retinene. These properties account for (1) the demonstrable shift in sensitivity to longer wavelengths in going from rod to cone vision in a variety of animals (the Purkinje shift) and (2) the obviously faster recovery of cone sensitivity which occurs during dark adaptation. A fourth pigment, the existence of which is indicated by spectral sensitivity data, may be synthesized by incubating retinene<sub>2</sub> with cone opsin. The pigment which results has an absorption maximum at 620  $m\mu$  and is called cyanopsin. More subtle differences in the protein moiety of visual pigments may have extensive effects upon their properties. Within the rhodopsin class, for example, the range of absorption maxima is from 302  $m\mu$  (frog) to 478  $m\mu$  (some deep-sea fish). The rhodopsins of certain fish and of the alligator, moreover, may have regeneration rates radically faster than those of cattle and frogs.

It is clear that cone pigments exist which have not as yet been extracted. It has been regarded as essential that animals with trichromatic color vision have cones with at least three different spectral sensitivities; and recently, by a technique of reflection photometry, the existence of three cone pigments in the normal human eye has been demonstrated. The absence of specific ones, moreover, has been shown to correlate with the major types of color blindness.

**Invertebrates.** The visual pigments of invertebrates have not been extensively studied, but those identified to date appear to be chemically similar to those of vertebrates. The squid has a rhodopsin based on retinene<sub>1</sub>, with the interesting difference that, when the pigment is photoexcited, the retinene moiety isomerizes but remains attached to its site in the protein instead of hydrolyzing away as in vertebrate rhodopsins. Photosensitive pigments based upon retinene have also been recently isolated from the eyes of crustaceans and insects.

**Electrical activity of photoreceptors.** Despite the extensive information about the chemistry of visual pigments, the mechanism by which their photochemical activation is translated into nerve impulses remains obscure.

**Vertebrates.** It has been known since the late 1800s that the vertebrate retina generates an electrical potential upon illumination. The record of this potential, called the electroretinogram (ERG), is polyphasic in wave form, with an initial cornea-negative component (the *a* wave), a large, cornea-positive component of longer latency (the *b* wave), and an off-response. Although it has been felt by some that the ERG is a receptor potential produced by the rods and cones, it is probable that the neural layers of the retina, the bipolar and ganglion cells, generate much of it. Nevertheless, the ERG is an accurate index of visual sensitivity; it has been shown that human dark adaptation and spectral sensitivity curves determined electroretinographically agree well with those obtained using subjective

measurements. The method is, therefore, useful for obtaining data on spectral sensitivity and dark adaptation in a variety of animals, and has often been employed in preference to behavioral techniques. The ERG has also proven useful clinically in the early diagnosis of diseases involving the retina.

Unfortunately, however, the uncertainty as to the source of the ERG has limited its use as a tool in the analysis of primary visual events. For this purpose, the technique of single-unit recording employing small electrodes is the method of choice. This method, however, has not been successfully applied so far to primary sensory cells in the vertebrate retina. A considerable body of information concerning the responses of retinal ganglion cells exists, but the ganglion cell is two synaptic steps removed from the primary event of photoreception. Therefore, such studies have been primarily directed toward problems of neural organization in the retina rather than toward the problem of photoexcitation.

**Invertebrates.** Invertebrate photoreceptors, especially the compound eye of *Limulus*, the horseshoe crab, have afforded an opportunity for studying the responses of nerve fibers from presumed primary photoreceptor cells. Experiments of this sort, in which impulse discharges from a single primary nerve fiber are studied in response to various intensities and durations of illumination delivered to the sense cell have demonstrated (1) that the frequency of impulse discharge, up to some maximum, is linearly related to the logarithm of the stimulus intensity, (2) that within a certain critical duration, usually somewhere between 0.1 and 1.0 sec, stimuli in which the product of intensity and duration is constant produce discharges of identical frequency, (3) that the latency of the impulse discharge decreases as the intensity is increased. Some similar conclusions had been reached earlier on the basis of behavioral studies. In addition, these electrophysiological recording techniques from primary sensory fibers have been used to measure spectral sensitivity and dark adaptation.

The electrical response to illumination in the *Limulus* ommatidium has also been recorded inside single sense cells, using micropipette electrodes. Electrical responses recorded in this way fall into two classes: (1) in the majority of cases, the response to illumination is a sustained depolarization on which small spikes, or none, are superimposed; (2) more rarely, the response consists of a similar depolarization upon which large spikes are superimposed. The depolarization is a light-induced generator potential which apparently initiates the nerve impulse discharge in the axon. The frequency of the impulse train discharged is, within limits, directly proportional to the amplitude of the generator potential, and the generator potential amplitude is proportional to the logarithm of the light intensity.

**Light and depolarization.** It may thus be generally true that light energy, absorbed by a photo-

sensitive pigment, acts to produce a sustained depolarization of sense cells by increasing the permeability of the cell membrane, and that this depolarization acts as a generator potential to initiate nerve impulses. The nature of the mechanisms by which the absorption of light and the chemical changes which follow it trigger this generator potential is, however, still a mystery.

[D. KENNEDY; V. J. WULFF]

## Photosphere

The photosphere of the Sun (the visible surface of the Sun or other stars) is a gaseous layer a few hundred kilometers thick with an average effective temperature of 5780°K, determined from the total radiation per square centimeter. The temperature is maintained by convection which brings hot material from the opaque solar interior to the surface in the form of rising columns of gas (see STELLAR EVOLUTION). The convection produces a small-scale granular texture which is visible through a projection telescope. In areas where strong magnetic fields inhibit the convection, the photosphere cools and dark sunspots appear. See SUN.

[J. W. EVANS]

## Photosynthesis

Literally, synthesis of chemical compounds in light. The term photosynthesis, however, is used almost exclusively to designate one particularly important natural process of this type: the manufacture, in light, of organic compounds (primarily certain carbohydrates) from inorganic materials, with simultaneous liberation of oxygen, by chlorophyll-containing plant cells. This process requires a supply of energy in the form of light, since its products (carbohydrates and oxygen) contain much more chemical energy than its raw materials (water and carbon dioxide). This is clearly shown by the liberation of energy in the reverse process, the combustion of organic material with oxygen. See PLANT RESPIRATION.

The light energy, taken up by the pigments of the photosynthesizing cells, especially by the green pigment, chlorophyll, is partially converted by photosynthesis into stored chemical energy. Together, the two aspects of photosynthesis—the conversion of inorganic into organic matter, and the conversion of light energy into chemical energy—make it the fundamental process of life on earth. It is the unique source of all living matter and of all life energy on our planet.

Since fossil fuels (coal, oil, peat) are half-decayed products of plant photosynthesis from past geological ages, it can be said that not only all life energy, but also nearly all industrial power, as well as all domestic heat, have their origin in photosynthesis. Exceptions are wind and water power and nuclear power.

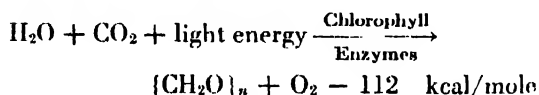
The elements carbon, oxygen, and hydrogen are exchanged, through photosynthesis and respiration, in an endless cycle, between the organic and the inorganic worlds. However, one ingredient of photosynthesis—light energy—is not regenerated in

this cycle. Therefore, life on earth can be maintained only by the constant supply of solar energy and its utilization through plant photosynthesis.

**The speed of photosynthesis.** Under favorable external conditions, photosynthesis is a remarkably fast process. With an adequate supply of carbon dioxide and light, a green cell will produce as much as 30 times its own volume in oxygen every hour. The rate of photosynthesis can be varied by varying the supply of carbon dioxide, the intensity or color of illumination, or the temperature. In addition to these easily controllable external conditions, the rate of photosynthesis depends also on the age, nutrition, and physiological condition, within the organism, factors which are much more difficult to define and control precisely.

**Turnover of photosynthesis on earth.** The total turnover of photosynthesis on earth has been roughly estimated in two ways: by averaging the yields of organic matter per unit area of field, forest, steppe, and ocean; and by determining the average utilization of incident solar energy by vegetation-covered areas (which is of the order of 1% if the whole solar spectrum is taken into consideration, or 2% if only visible light is considered). Both procedures lead to numbers of the magnitude of 100,000,000,000 ( $10^{11}$ ) tons of carbon transferred from the inorganic into the organic state each year. This corresponds to about  $10^{11}$  kilocalories ( $10^{15}$  kilowatt-hours) of light energy stored annually. The estimate is rough, mainly because of uncertainty as to the average rate of photosynthesis in the world's oceans.

**The over-all reaction.** The net over-all chemical reaction of photosynthesis is



where  $\{\text{CH}_2\text{O}\}_n$  stands for a carbohydrate (sugar). All oxygen liberated in photosynthesis originates in water, and none in carbon dioxide, as shown by experiments with isotopic tracers in which tracer oxygen was found in liberated  $\text{O}_2$  when it was incorporated into water but not into carbon dioxide. These experiments were made by S. Ruben, M. Randall, M. D. Kamen, and J. Hyde in 1941.

The photochemical reaction in photosynthesis belongs to the type known as oxidation-reduction, with carbon dioxide acting as the oxidant (hydrogen or electron acceptor) and water as the reductant (hydrogen or electron donor). The unique characteristic of this particular oxidation-reduction is that it goes "in the wrong direction," converting chemically stable materials into chemically unstable products. Light energy is used to make this "uphill" reaction possible, and a considerable part of the light energy utilized is stored as chemical energy [112 kilocalories per mole, or 44 grams, of reduced carbon dioxide, as indicated in Eq. (1)].

**A multistage process.** From an enormous amount of research by plant physiologists, biochemists, photochemists, and biophysicists, it is



known that photosynthesis is a complex, multistage process. Its main parts are the primary photochemical process, in which light energy, taken up by chlorophyll, is converted into chemical energy, in the form of some energy-rich intermediate products, and enzyme-catalyzed "dark" (that is, not photochemical) reactions, by which these intermediates are converted into the final products—carbohydrates and free oxygen. These reactions of photosynthesis can be grouped into three phases, as shown in the scheme of Fig. 1. Phase 1 is the evolution of oxygen from dehydrogenated water by a series of dark reactions. This is the least-known aspect of photosynthesis. About all that is known is that it is enzymatic and requires manganous ions. Phase 2 is the transfer of hydrogen atoms, H (or electrons)—not of hydrogen molecules,  $H_2$ —from an unknown intermediate in phase 1 to some intermediate acceptor capable of reducing carbon dioxide. This is the light phase of photosynthesis. Phase 3 is the reduction of carbon dioxide by a series of dark reactions. The use of radioactive carbon (carbon 14) as a tracer has given considerable insight into the nature of these reactions. This phase, like phase 1, occurs at a more or less constant level of energy.

**The Hill reaction.** Various observations suggest that the immediate action of light (the primary photochemical process) in photosynthesis involves the transfer of hydrogen (or electrons) from water or from a large molecule into which water has been incorporated in a preparatory step) to an acceptor (primary oxidant) X. The latter is not yet firmly identified molecule. Some believe it to be nicotinamide adenine dinucleotide phosphate, usually abbreviated as NADP (previously called TPN); but it could also be another biological catalyst, such as ferredoxin, with a similarly high, or even higher, reduction potential—that is, a compound which is relatively unstable in the reduced form, and thus has a strong tendency to transfer its hydrogen to other molecules.

These conclusions are made plausible by consideration of the so-called Hill reaction (named after its discoverer, R. Hill). This reaction is a process in which illuminated chlorophyll-bearing fragments of plant cells produce oxygen from water, without concomitant reduction of carbon dioxide, but with the reduction of added, less stable oxidants, such as quinone or ferricyanide. Since the quantum requirement and other kinetic characteristics of the Hill reaction prove to be similar to those of photosynthesis, it can be assumed that in the Hill reaction the primary photochemical apparatus of photosynthesis is preserved more or less intact. In this reaction, however, the coupling of the primary photochemical process with the enzymatic mechanism which brings about the reduction of carbon dioxide is impaired by the mechanical destruction of the cell.

In 1954, with the use of  $C^{14}$  as a radioactive tracer, it was observed that certain organic compounds, containing the tracer and having reduction levels up to that of sugars, are formed by illumina-

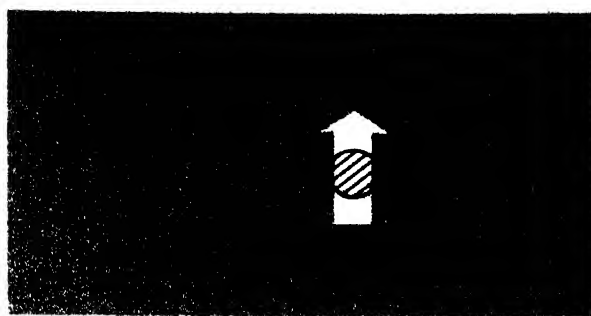


Fig. 1. Schematic illustration of photosynthesis. Phase 1, oxidation of water, consists of enzymatic reactions converting dehydrogenated water to free oxygen. Phase 2, the light reaction, is the transfer by light-excited chlorophyll (Chl) of hydrogen (or electrons). Phase 3, reduction of carbon dioxide, consists of enzymatic reactions converting carbon dioxide and light-supplied hydrogen to carbohydrates ( $CH_2O_n$ ).

nating whole or fragmented chloroplasts in the presence of  $C^{14}$ -labeled carbon dioxide, provided certain auxiliary substances are supplied. This suggests that the coupling of the photochemical apparatus with the carbon dioxide enzymatic system is not entirely lost by the mechanical destruction of the cells; or, at least, that this coupling can be partially restored by the addition of these compounds.

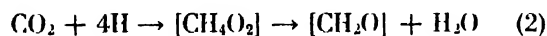
**The quantum process of photosynthesis.** In photosynthesis, the energy of light quanta is converted into chemical energy. In the conversion of 1 mole of  $CO_2$  and 1 mole of  $H_2O$  into 1 mole of carbohydrate groups,  $CH_2O$ , and 1 mole of oxygen, according to Eq. (1), about 112 kilocalories of total energy, or, under natural conditions, about 120 kilocalories of potential chemical energy ("free energy") are stored. Light is absorbed by matter in the form of quanta of energy (photons). A 2% energy conversion yield means that an average of considerably over 100 quanta are absorbed by the pigments, under natural conditions, to bring about the reduction of one molecule of carbon dioxide. See ABSORPTION (ELECTROMAGNETIC RADIATION); PHOTON.

Under natural conditions, carbon dioxide supply is not always adequate, while light supply may be overabundant for most effective utilization. Furthermore, not all plant cells are in the most productive physiological state. By using turbulently flowing suspensions of microscopic unicellular algae in carbon dioxide-enriched water, a utilization of up to 7% of absorbed visible sunlight has been obtained in large-scale experiments. Under still more favorable small-scale laboratory conditions (very weak illumination and very effective carbon dioxide supply to the algae by strongly stirred carbonate-bicarbonate buffer solutions), up to 30% of absorbed light energy could be converted into stored chemical energy, corresponding to a quantum requirement of about 8 quanta per molecule of carbon dioxide. (These measurements were made by



R. Emerson and several other investigators.) This is a very high efficiency, not matched by any known petrochemical reaction.

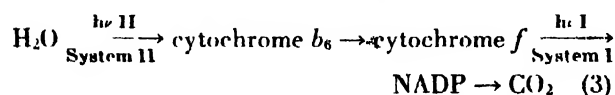
It is not impossible, however, that under some conditions, not yet clearly established, quantum requirements of less than 8 can be obtained, perhaps 7 or even 6, although as yet no such results have found general acceptance by workers in the field. The answer is of considerable importance for the interpretation of the mechanism of light action in photosynthesis. The reduction of one molecule of carbon dioxide to the carbohydrate level requires the transfer of four hydrogen atoms:



A quantum requirement of 8 (or more) would thus permit two quanta to be used for the transfer of each hydrogen atom (or electron).

**The two-quanta hypothesis.** A specific mechanism, in which two quanta are used to transfer one hydrogen atom in photosynthesis, was suggested by experiments of Robert Emerson in 1956-58. Earlier (1943), Emerson had discovered that the "quantum yield of photosynthesis" (number of  $\text{O}_2$  molecules evolved per absorbed quantum), while constant at the shorter wavelengths of light (red, orange, yellow, green), declines in the dark red above 680  $\text{m}\mu$  (the "red drop"). Thirteen years later, Emerson found that this low yield can be enhanced if both chlorophyll *a* and *b* are simultaneously excited (only chlorophyll *a* absorbs above 680  $\text{m}\mu$ ). The "Emerson enhancement effect" suggested that two pigments must be excited to perform efficient photosynthesis and thus indicated involvement of two light reactions in photosynthesis, one sensitized by light absorption in chlorophyll *a*, and one by absorption in another pigment (for example, chlorophyll *b*).

R. Hill and F. Bendell proposed, in 1960, that one of these reactions is the transfer of hydrogen (or of an electron) from some intermediate in the conversion of water to oxygen to a cytochrome (specifically, so-called cytochrome  $b_6$ ), while the other is the transfer of hydrogen (or electron) from another cytochrome (specifically, cytochrome *f*) to an intermediate in the conversion of carbon dioxide to carbohydrate. The intermediate transfer of hydrogen (or electron) from cytochrome  $b_6$  to cytochrome *f* can occur by a dark reaction, because the former is a stronger reductant than the latter. The "bucket brigade" for the transfer of hydrogen can thus be represented as follows:



where  $h\nu$  represents a light quantum.

The designation of the first light reaction in the sequence as II, and of the second one as I, is somewhat confusing but is widely used. Experimental evidence for the key role of cytochromes in this sequence was provided in 1960 by L. N. M. Duysens.

Whether two cytochromes are involved, as suggested in Eq. (3), or just one, and whether NADP or ferredoxin or some other compound is the immediate acceptor of hydrogen in light reaction I remains uncertain. Also uncertain as yet is the possible role of other experimentally identified oxidation-reduction catalysts such as plastoquinone, plastocyanin, and certain flavins whose reductor potentials are close to those of the cytochromes  $b_6$  and *f*.

Subsequent experiments by Govindjee and E. I. Rabinowitch, C. S. French, L. R. Blinks, J. Myers, H. Giffon, D. Fork, L. N. M. Duysens, J. B. Thomas, C. P. Whittingham, and others enlarged Emerson's observation by suggesting that plants contained two pigment "systems." One (system I, sensitizing reaction I) contains the major part of chlorophyll *a* the other (system II, sensitizing reaction II) contains chlorophyll *b* and other auxiliary pigment (for example, the red and blue pigments, called phycobillins, in red and blue-green algae, and the brown pigment fucoxanthol in brown algae and diatoms). The data of Govindjee and E. I. Rabinowitch suggest that system II contains also some chlorophyll *a* (in green cells and diatoms it is a special form of it with an absorption band at 670  $\text{m}\mu$ , chlorophyll *a* 670). It appears that efficient photosynthesis requires the absorption of an equal number of quanta in system I and in system II, and that within both systems excitation energy undergoes resonance migration from one pigment to another, until it ends in a special molecule of chlorophyll *a*, the latter then enters into the chemical reactions.

The resulting scheme of photosynthesis is shown in Fig. 2. It includes the recent findings of B. Kok, H. W. Wilt, L. N. M. Duysens, D. I. Arnon and several others. The two vertical arrows suggest the two light reactions (transfers of hydrogen atoms or electrons), the horizontal or slanted arrows the dark reactions. On the left of this diagram is a scale of oxidation-reduction potentials,  $\epsilon_0$ .

**Photophosphorylation.** It has been shown on bacterial material by A. Frenkel and on chloroplast material by D. I. Arnon and coworkers that green plants that these pigment-bearing particles, when illuminated in the presence of ADP (adenosine diphosphate) and inorganic phosphate, use light energy to synthesize ATP (storing about 10 kcal of converted light energy in each molecule of the high-energy phosphate ATP). This photophosphorylation could be associated with some energy releasing step in photosynthesis. A possible location of this step is shown in Fig. 2—the reduction of cytochrome *f* by reduced cytochrome  $b_6$ . This would be analogous to the way in which ATP is produced in respiration. The ATP produced in the light stage of photosynthesis apparently is needed to make some later, enzymatic reactions (such as the reduction of a carboxylic acid by reduced NADP) run in the needed "uphill" direction. See ADENOSINEDIPHOSPHATE (ADP); ADENOSINETRIPHOSPHATE (ATP).

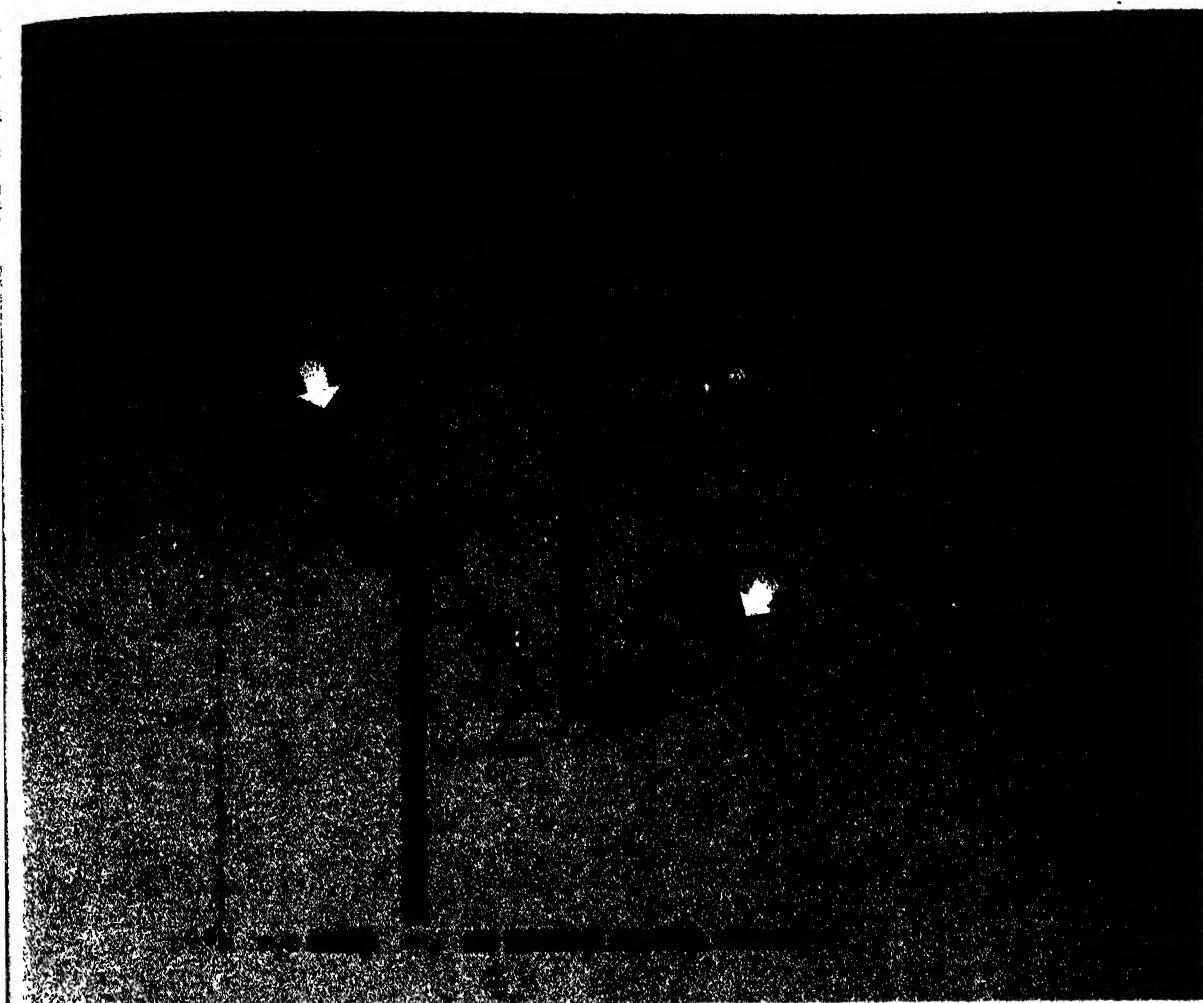


Fig. 2. Two light steps in photosynthesis (compare with Fig. 1).  $ZH_2$  is the (unknown) primary donor of hydrogen atoms (or electrons). ADP is adenosine diphosphate and ATP is adenosine triphosphate (high-energy phosphate). P680 is an unknown "trap" in pigment system II, P700 is pigment 700, the energy trap for system I. NADP is nicotinamide adenine dinucleotide phosphate, FD is ferredoxin, FP is flavoprotein, PGA is phosphoglyceric acid, RuMP is ribulose monophosphate, and RuDP is ribulose diphosphate. The two photosynthetic units involved in the

process are indicated by dots. Unit I contains mainly long-wave forms of chlorophyll *a*, with (in the case of red and blue-green algae) a certain amount of phycoerythrin and phycocyanin. Unit II contains chlorophyll *a* 670 and all "accessory" pigments—chlorophyll *b* in green cells, phycoerythrin and phycocyanin in red and blue-green algae, fucoxanthol and chlorophyll *c* in brown algae. The light energy absorbed in I is delivered, by transfer, to a molecule of P700; by analogy, the energy absorbed in II is supposed to be delivered to a molecule of the hypothetical pigment P680.

**Saturation: light and dark reactions.** If the rate of photosynthesis is plotted as a function of light intensity, a curve results which shows first a proportional increase, then a gradual saturation. This saturation could be due to various causes. One is the limitation of carbon dioxide supply from the outside. Further increase of light intensity becomes of no use when all carbon dioxide molecules reaching the cell are used up as fast as they arrive. Carbon dioxide concentration can thus act as a limiting factor. (The same principle applies to the effect of increasing carbon dioxide concentration in weak light, when the reaction is light-limited.)

The concept of limiting factors was introduced

by F. Blackmann into photosynthesis in 1905. It is not a special characteristic of photosynthesis but applies to all chemical systems in which one or several reactants must be continuously supplied from the outside to keep the reaction going. (Light can be considered as a reactant in photochemistry.)

When the supply conditions for carbon dioxide and light are most favorable, the rate of photosynthesis still shows saturation. This is generally attributed to the need, for the completion of photosynthesis, of at least one (and more likely, several) light-independent enzymatic reactions. An enzyme-catalyzed reaction has a certain maximum rate,  $E_0/t$ , determined by the total amount of its catalyst

(enzyme) available in the cell,  $E_0$ , and its turnover time,  $t$ , which is the average time the enzyme molecule must work (at a given temperature) on a molecule of the reaction substrate before its transformation is completed:

$$V_{\max} = E_0/t \quad (4)$$

The several enzymes involved in photosynthesis thus impose ceilings on the maximum speed at which photosynthesis as a whole can proceed, each enzyme functioning as a bottleneck of limited capacity in the reaction path. The enzyme which imposes the most effective (lowest) ceiling seems to be involved in the liberation of oxygen rather than in the reduction of carbon dioxide, since the same saturation rate is observed also in the Hill reaction.

**Photosynthesis in flashing light.** In 1932 it was shown by R. Emerson and W. Arnold how the light reaction in photosynthesis can be separated from the dark reaction by the use of brief, intense light flashes, separated by intervals of darkness of variable duration. The main conclusion was that the maximum yield of  $O_2$  from a single flash is about one molecule of oxygen per 2500 molecules of chlorophyll present in the cell. This can be interpreted as meaning that the cells contain one molecule of the rate-limiting enzyme per 2500 chlorophyll molecules; but if the same enzyme has to work  $n$  times for the liberation of one molecule of  $O_2$ , this ratio must be reduced to  $2500/n$ ; for example, if  $n = 8$ , it becomes about 300.

The use of flashing light, with varying flash intensity and duration, variable flash grouping, and varying dark intervals, is one of the most important approaches to the understanding of the way in which different factors affect the over-all rate of photosynthesis through their effects on different steps in the reaction sequence. Monochromatic flashes have been used to gain understanding of the mechanism of the two light reactions. The experiments of C. S. French show that the product of light reaction I has a half-life of 18 seconds (in red algae), while that of reaction II seems short-lived.

**Photochemical apparatus.** The primary photochemical stage of the photosynthetic process appears to be closely associated with certain structural elements found in plant cells. All algae (except the primitive blue-green algae), as well as all higher plants, contain pigment-bearing intercellular bodies called chloroplasts. In the leaves of the higher land plants, these are usually flat ellipsoids about 5000  $m\mu$  (0.005 mm) in diameter and 2000–3000  $m\mu$  in thickness; 10–100 of them may be present in an average cell of leaf parenchyma. See LEAF BOTANY.

In algae the number and shape of chloroplasts are much more variable; for example, the much-studied green unicellular alga *Chlorella* contains only one bell-shaped chloroplast.

All chloroplasts fixed (solidified, usually by means of osmic acid) and sliced show under the



Fig. 3. Electron micrograph of a cross section of a chloroplast of corn (*Zea mays*), fixed with osmic acid, and sliced. This micrograph shows lamellae and cylindrical grana formed by their local reinforcement. Chloroplasts of some other species show only lamellae, no grana. (After A. Vatter)

electron microscope a layered structure, with alternate lighter and darker layers roughly 10  $m\mu$  in thickness. It is generally assumed that these layers differ in the proportion of proteidic and lipidic (fatlike) substances in them.

Two main types of chloroplasts are known. In some, the layered structure extends more or less uniformly through the whole chloroplast body (lamellar chloroplasts). In others, this structure is emphasized in certain cylindrical sections and is less pronounced between them (Fig. 3). When such granular chloroplasts are permitted to dry out and disintegrate, stacks of discs break out of the structure, and appear as cylindrical grana in the electron micrograph.

In photosynthesizing bacteria and in the lowest truly photosynthetic plants (blue-green alga Cyanophyta), the photochemical apparatus is more primitive. However, lamellae similar to those in chloroplasts have been observed also in blue-green algae (and much smaller lamellar particles, in bacteria). The true unit of photochemical apparatus may be a lamella consisting of two membranes. Evidence of a "cobblestone" appearance of the lamellae in higher plants has been noted on electron micrographs. These "cobblestones" could be perhaps identified with photosynthetic units (containing about 300 chlorophyll  $a$  molecules); the way in which these units are attached to the membranes remains uncertain. The "cobblestones" have been given the name *quintasomes* to indicate that they are the probable sites of the light reaction in photosynthesis (Fig. 4).

**Distribution of chlorophyll.** It is generally as-

assumed that chlorophyll molecules are located at the interfaces between the proteidic and the lipoidic layers of the chloroplasts, perhaps forming one-molecule-thick cohesive layers (monolayers). Estimates suggest that the total area of such interfaces in a chloroplast is just about adequate to accommodate all the chlorophyll molecules present, allowing about  $1 \text{ m}\mu^2$  for each molecule.

What could be the purpose of a laminar structure "painted over," as it were, with monolayers of pigments? Two hypotheses are offered, both of which may be correct. One is that the two-dimensional, laminar structure creates the best conditions for easy access of the reaction substrates to the chlorophyll molecules, and also for rapid removal of the reaction products. Considering that photosynthesis is by far the fastest metabolic process in the cell, easy supply and removal of reactants may be an important advantage. (In terms of number of molecules transformed per unit volume per unit time, photosynthesis can be 10 or 20 times faster than respiration.) The other hypothesis places emphasis on the possibility of excitation-energy migration in the pigment layer. The absorption of a photon activates a single chlorophyll molecule, transferring it into a short-lived, energy-rich excited state. To minimize the danger of the dissipation of this excitation energy before it can be used for photochemical purposes, it may be advantageous to permit the energy to move around, jumping from one chlorophyll molecule to another, thus increasing the chance of its encounter with the reaction substrates. Such a mechanism of resonance energy migration is in fact postulated in some theories of the primary photoprocess in photosynthesis. It is a

plausible, but by no means a proven, concept.

However, the picture of chlorophyll molecules distributed in uniform monolayers on interfaces between proteinaceous and lipoidic lamellae may be oversimplified. There is considerable evidence that not all chlorophyll *a* molecules in the cell are in the same state. These kinds of chlorophyll differ in the positions of their absorption bands and in their capacity for fluorescence, and they may have different functions in photosynthesis. Only a small fraction of chlorophyll *a* may be closely associated with the primary photochemical process, while the rest serves primarily as energy-supplier to it. These differences in function must be somehow associated with the spatial arrangement of chlorophyll molecules in the layered structure, but just how is as yet unknown.

**Accessory pigments.** An interesting problem is also the location in the chloroplasts, and the function in photosynthesis, of so-called accessory pigments—that is, pigments other than chlorophyll *a*, the one pigment present in all photosynthetically active plants. In the first place, there are other chlorophylls, such as chlorophyll *b* in higher plants and green algae, and chlorophyll *c* in brown algae. Then there are nonchlorophyllous pigments, belonging to two groups: (1) The carotenoids, so called because of similarity to the orange pigment of carrots, are a variable assortment of pigments found in all photosynthetic higher plants, algae, and even bacteria. (2) The phycobilins, or "vegetable bile pigments," are chemically related to animal bile pigments. The phycobilins are either red (phycoerythrins) or blue (phycocyanins). Both types are present in red algae (*Rhodophyta*) and blue-green algae (*Cyanophyta*), the red pigment prevailing in the first group of organisms and the blue pigment prevailing in the second group. See CAROTENOID.

In 1884, Engelmann suggested that all these pigments contribute to photosynthesis. Later it was concluded that only light absorbed by chlorophyll was of importance. However, it is now clear that light absorbed by accessory pigments does contribute to photosynthesis. These conclusions are derived from measurements of the so-called action spectra of photosynthesis obtained primarily by R. Emerson, and by F. T. Haxo and L. R. Blinks. In such measurements photosynthesis is produced by monochromatic light, isolated by means of a spectrophotometer, and the production of oxygen (or consumption of carbon dioxide) per absorbed quantum of light (the quantum yield) is measured as a function of wavelength. The observed spectral variations in the quantum yield of photosynthesis can be related to the proportions of light absorbed, at each wavelength, by the different pigments in the cells. Measurements of this kind led to the conclusion that quanta absorbed by carotenoids are 50–80% less effective than those absorbed by chlorophyll *a* in contributing energy to photosynthesis. An exception is fucoxanthol, the carote-

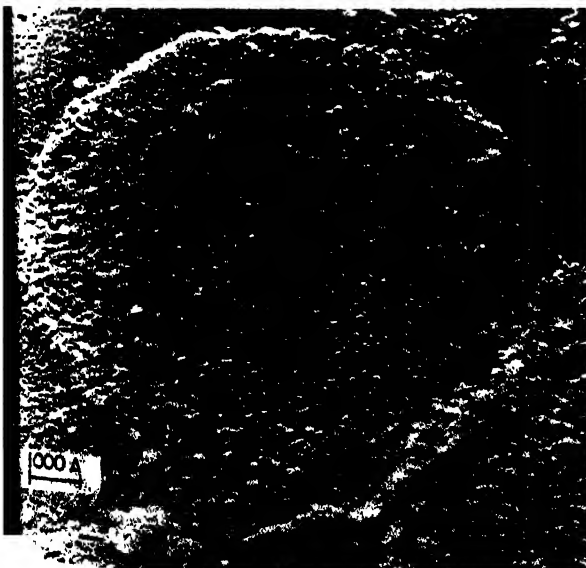


Fig. 4. Membranes containing chlorophyll taken from spinach chloroplast. This chromium-shadowed preparation shows that the membrane is composed of a highly ordered array of units called quantasomes. ( $\times 187,000$ ) [After R. B. Park, courtesy of Science, 144 (1962), 1964]

noid that accounts for the color of brown algae (*Phaeophyta*) and diatoms (*Bacillariophyta*). This pigment supplies light energy to photosynthesis about as effectively as the green pigment. The red and blue pigments of the *Rhodophyta* and *Cyanophyta* are also highly effective, as effective as chlorophyll or somewhat less, depending, among other things, on the history of the algae, particularly the color of the light to which they have become adapted.

#### **Energy transfer between pigment molecules.**

Chlorophyll *a* in plant cells is weakly fluorescent; this means that some of the light quanta absorbed by it (up to 3%) are reemitted as light. Observations of the action spectrum of chlorophyll fluorescence in different plants have suggested close parallelism with the action spectrum of photosynthesis. In other words, fluorescence of a form of chlorophyll *a* in the plant can be excited also by light absorbed by the accessory pigments, with the probability of this sensitized fluorescence closely paralleling that with which the same light is used for photosynthesis. (These measurements were made by H. J. Dutton and coworkers, L. N. M. Duysens, C. S. French, and others.) Excitation of chlorophyll fluorescence by light quanta absorbed by phycoerythrin requires transfer of the excitation energy quantum from the primarily excited phycoerythrin molecule to a nearby chlorophyll molecule (as in acoustic resonance, where striking one bell causes another nearby bell to ring). Therefore, it can be suggested that light quanta absorbed by accessory pigments, such as carotenoids and phycobilins, contribute to photosynthesis by being transferred to chlorophyll *a*. By this mechanism, red algae, growing relatively deep under the sea, where only blue-green light penetrates, can supply the energy of this light to chlorophyll, which does not itself absorb it.

If excitation energy can be transferred efficiently, in the chloroplasts, from accessory pigments to chlorophyll, there is a good probability that a similar transfer can and does occur also between different chlorophyll molecules themselves. If this happens repeatedly during the lifetime of excitation, the excitation energy can migrate as much over considerable distances in the chloroplast. As suggested in the section on distribution of chlorophyll, this migration of excitation energy may have advantages from the point of view of efficient utilization of absorbed light quanta for photosynthesis.

**Electron transfer in chloroplasts.** It has also been suggested that absorption of a light quantum in the dense layer of chlorophyll molecules may lift an electron into a state in which it will be able to move through the lamella. This is comparable to photoconductivity, a phenomenon known to occur in certain insulating crystals, which become electric conductors when irradiated with light. In this way, an electron may become spatially separated from the positive chlorophyll ion and may then act as a reductant at some location in the chloroplast

structure (addition of an electron is equivalent to reduction—compare, for example, the conversion of ferric ion,  $\text{Fe}^{+++}$ , to ferrous ion,  $\text{Fe}^{++}$ ), while the positive ion may act as an oxidant by taking an electron away from a substrate in another place. Thus the oxidation and the reduction products of the light reaction will be spatially separated, and the danger of their recombination, with the loss of stored energy, reduced.

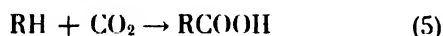
This picture of photosynthesis as a process typical of a solid, crystalline medium, rather than of a solution, is a tempting one; it has been supported by certain experiments on chloroplast films. However, other considerations speak against it, such as the similarity of the shape of the absorption band of chlorophyll in the living cell with its shape in solution, and the fact that electrons cannot remain free in the presence of water. Perhaps the solid-state theory applies only to very small regions in the chloroplasts or grana, containing 10 or 100 pigment molecules.

The concept of electron transfer is used, in the theory of photosynthesis, also in a somewhat different sense. In the section on the two-quanta hypothesis, the scheme of photosynthesis was discussed in which two cytochromes were involved as intermediates between the two photochemical steps. Cytochromes are known to be oxidized by conversion of their iron atoms from the  $\text{Fe}^{++}$  to the  $\text{Fe}^{+++}$  state, by loss of an electron. The intermediate enzymatic state in photosynthesis thus represents a “downhill” electron transfer, similar to that in respiration. At the two ends, however, the oxidation of water and the reduction of carbon dioxide must involve the loss and the acquisition of hydrogen atoms—that is, of electrons and protons. It remains uncertain where in the sequence of reactions the hydrogen transfer is replaced by electron transfer, and again by hydrogen transfer—in particular, whether the two primary photochemical reactions themselves result in the one or the other.

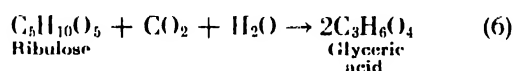
**Chemical role of chlorophyll.** Unless a solid-state picture of the primary photochemical process in photosynthesis is assumed, the question arises: How does the chlorophyll *a* molecule, ultimately in possession of the absorbed quantum of energy, utilize it for an energy-storing photochemical process, such as the transfer of a hydrogen atom from a reluctant donor (perhaps water) to a reluctant acceptor (perhaps nicotinamide adenine dinucleotide phosphate, NADP)? It has been suggested that chlorophyll acts as a typical oxidation-reduction catalyst—that is, by being itself first oxidized and then reduced, or vice versa, with the difference that it uses its excitation energy, either in one or in the other, or in both, of these steps. Support for this plausible hypothesis is provided by observations of reversible photochemical oxidation and of reversible photochemical reduction of chlorophyll in solution. Studies of changes in the absorption spectrum of photosynthetic cells in light suggest that a small fraction of a special form of chlorophyll *a*, absorbing maximally at 700 and

30 m $\mu$ , finds itself, during illumination, in a changed state.

**Carbon dioxide reduction.** Since 1939, knowledge of the conversion of carbon dioxide into organic molecules, such as glucose, has been much advanced by the application of radioactive tracers, particularly of C<sup>14</sup> by M. Calvin and coworkers. It has been long assumed that the molecule CO<sub>2</sub> is not reduced photochemically as such but is first incorporated into a larger molecule. The process is now generally assumed to occur by way of carboxylation, that is, formation of an organic acid from hydrogen-containing organic molecule:



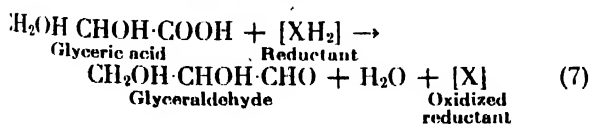
By the use of C<sup>14</sup> tracer, it has been found that the compound RH is a pentose, that is, a sugar with only five carbon atoms, instead of the six present in the more common hexoses. The pentose involved is called ribulose, more precisely, a phosphate ester of this sugar, ribulose diphosphate. It is still uncertain whether the carboxylation of this compound is normally accompanied by hydrolytic splitting, giving rise to two molecules of phosphoglyceric acid, as indicated in the following equation:



For simplicity the phosphate residues are omitted in Eq. (6).

The alternative to Eq. (6) is the formation of a single molecule of an acid with a six-membered carbon chain. Phosphoglyceric acid has been found by some workers to be the main C<sup>14</sup>-containing product after very brief (1–10 sec) photosynthesis of algae in C<sup>14</sup>-tagged carbonate. However, in these experiments, algae were killed, at the end of exposure, by dropping them into boiling alcohol, and it has been suggested that this may have caused the decomposition of a 6-carbon acid into two molecules of phosphoglyceric acid.

If phosphoglyceric acid is the normal intermediate in photosynthesis, then it is reasonable to postulate that the next step, after its formation, is its reduction to phosphoglyceraldehyde:



In this equation [XH<sub>2</sub>] stands for a reduced compound ready to give away two atoms of hydrogen, leaving the hydrogen carrier [X] behind. This strong reductant must be supplied by the primary photochemical process.

The pyridine nucleotide NADP, nicotinamide adenine dinucleotide phosphate, serves as a mediator between the primary photochemical oxidation-reduction and the enzymatic reduction of carbon dioxide in photosynthesis. It has been proved that

NADP can be reduced to NADPH<sub>2</sub> by illuminated chloroplast suspension in the Hill reaction; however, this is not in itself convincing proof of the postulated participation of this compound in photosynthesis, because many different oxidants can be reduced under the same conditions. However, it has been shown that NADP occurs in high enough concentration in chloroplasts to serve as a reductant in photosynthesis and that it does undergo photochemical changes in vivo.

One difficulty arises: NADPH<sub>2</sub> is not a strong enough reductant to reduce phosphoglyceric acid to phosphoglyceraldehyde (or, more generally, to reduce any carboxyl group, RCOOH, to the corresponding aldehyde, RCHO). In fact, the reverse reaction, oxidation of NADPH<sub>2</sub> by glyceraldehyde, liberates a considerable amount of energy. In respiration, this reaction is coupled with the conversion of adenosinediphosphate (ADP) and inorganic phosphate into adenosinetriphosphate (ATP), an energy-storing reaction in which the oxidation energy is neatly preserved in a so called high-energy phosphate bond, a widely used biological energy currency. It has been suggested that in photosynthesis the reverse happens—that is, the reduction of phosphoglyceric acid to phosphoglyceraldehyde by NADPH<sub>2</sub> is made possible by coupling it with the energy-supplying conversion of ATP back into ADP and inorganic phosphate.

This is the most common version of the mechanism of photosynthesis at present. Since glyceraldehyde has the reduction level of a carbohydrate (C<sub>n</sub>H<sub>2n</sub>O<sub>n</sub>, with H:O = 2:1), its enzymatic conversion to sugars, for example, to a hexose (as final product) or a pentose (as CO<sub>2</sub> acceptor, thus closing the cycle), can be accomplished without further need for light energy, by enzymatic reactions of the kinds well known from different metabolic pathways.

#### Bacterial photosynthesis and chemosynthesis.

Certain species of pigmented bacteria, some green (containing a green pigment called bacteriochlorophyll, or chlorobium-chlorophyll), some purple or red (containing bacteriochlorophyll and carotenoids), are able to synthesize organic matter from carbon dioxide in light. Since the main absorption band of bacteriochlorophyll is located in the near infrared while that of chlorophyll is in the red, purple bacteria can live also in invisible, infrared light. In contrast to green plants and algae, these organisms cannot use water as a source of hydrogen for the reduction of carbon dioxide, and can survive only under conditions providing other hydrogen sources, such as hydrogen sulfide or other sulfur compounds, free molecular hydrogen, or organic compounds. In the latter case, the bacteria destroy one kind of organic matter to synthesize another.

Because hydrogen is bound nowhere as strongly as in water, these types of photosynthesis store little, if any, light energy. They do not have the same significance as true photosynthesis in the transformation and storage of cosmic energy on



earth. In fact, all they can do is to utilize, in light, chemical energy already available in the form of unstable hydrogen compounds. In most cases light energy is used by them merely or mainly as chemical activation energy, as it is also used in most photochemical reactions in vitro.

It is uncertain whether bacterial photosynthesis involves one or two light reactions. Eight quanta seem to be required for the reduction of one  $\text{CO}_2$  molecule in bacteria but they show no "red drop" or Emerson effect. The absence of Emerson effect was shown by L. R. Blinks and C. B. Van Niel.

It is unknown whether bacterial photosynthesis is an earlier mode of life, preceding true photosynthesis, or a later form of life into which true photosynthesis has degenerated in chemical surroundings providing certain sources of hydrogen. In any case, bacterial photosynthesis is bound to remain limited to a few natural habitats, such as stagnant canal waters or volcanic sulfur springs.

For the sake of completeness, mention should be made also of chemosynthetic bacteria, cells which can achieve the conversion of carbon dioxide to organic matter with the help of hydrogen donors similar to those utilized by photosynthetic bacteria, but without the help of light. They simply burn chemical fuel by a mechanism permitting them to salvage some combustion energy to reduce carbon dioxide. In the simplest case, that of so-called hydrogen bacteria, the cells oxidize one part of molecular hydrogen to water and use some of the liberated energy to transfer another part of the hydrogen to carbon dioxide. Whereas photosynthetic bacteria can live anaerobically, the chemosynthetic ones require oxygen to keep their energy-liberating process in operation. Some chemosynthetic organisms have developed wherever oxidizable material is present in nature, be it coal, oil, free hydrogen, sulfur compounds, ammonia, nitrite, or ferrous salts. Again the question can be asked: What is the evolutionary role of the chemosynthetic way of life? Is it a predecessor of photosynthesis, or is it degradation of photosynthesis under especially "easy" conditions of abundant energy supply?

[GOVINDJEE; E. I. RABINOWITCH]

**Bibliography:** H. Gaffron, Energy storage in photosynthesis, in *Plant Physiology: a Treatise*, vol. IB, 1960; H. Gest et al. (eds.), *Bacterial Photosynthesis*, 1963; M. D. Kamen, *Primary Process in Photosynthesis*, 1963; B. Kok and A. T. Jagendorf (eds.), *Photosynthetic Mechanisms of Green Plants*, 1963; W. D. McElroy and B. Glass (eds.), *A Symposium on Light and Life*, 1960; E. I. Rabinowitch, *Photosynthesis and Related Processes*, 3 vols., 1945-1956.

## Phototransistor

A semiconductor device with electrical characteristics that are light-sensitive. Phototransistors differ from photodiodes in that the primary photoelectric current is multiplied internally in the device, thus increasing the sensitivity to light. For a discussion of this current multiplication property, see TRANSISTOR.

Some types of phototransistors are supplied with a third, or base, lead. This lead enables the phototransistor to be used as a switching, or bistable, device. The application of a small amount of light causes the device to switch from a low current to a high current condition. See PHOTOELECTRIC DEVICES. [W. R. SITNER]

## Phototube

An electron tube containing a photocathode from which electrons are emitted when it is exposed to light or other electromagnetic radiation. An elementary vacuum phototube consists of a photocathode, an anode, or electron collector, and an evacuated envelope through which radiation is transmitted to the photocathode. A gas phototube contains, in addition, an inert gas which may be ionized by electron current from the photocathode. For a description of a phototube in which the electron current is amplified by means of a secondary-emission electron multiplier, see PHOTOTUBE, MULTIPLIER.

Phototubes serve as sensing elements in the detection and measurement of light and ultraviolet or infrared radiation. Phototubes also convert variations in intensity of incident radiation into corresponding variations in electron output current, as in light-controlled relay circuits, and in the conversion of sound modulation of photographic film into audio-frequency currents, as in the sound tracks on motion-picture film.

The fundamental characteristics of a phototube are its spectral sensitivity characteristic, or output current expressed as a function of the wavelength of incident radiation at constant anode voltage, and its anode-current characteristics, which show the dependence of anode current on applied voltage and radiant flux input. Gas phototubes do not differ from vacuum phototubes with regard to spectral sensitivity characteristics, which are described below, but their anode-current characteristics are essentially different.

**Principles of operation.** The anode current of a vacuum phototube is directly proportional to the intensity of radiation incident on its photocathode. The anode is normally connected to a positive potential of at least 20 volts relative to the photocathode in order that the anode current be limited by photoelectric emission rather than by space charge or electron emission velocity.

In a gas phototube the photoelectric current is amplified by partial ionization of a gas contained in the tube at low pressure. An inert gas such as neon or argon is used because photocathodes react chemically with other gases. At low anode voltages the anode current of a gas phototube is emission-limited like that of a vacuum phototube. At anode potentials greater than 25 volts, electrons emitted from the photocathode acquire sufficient energy to ionize some of the gas atoms. The total current is then the sum of the free-electron current, the positive-ion current, and the current due to secondary electrons which are produced by ion bombardment of the photocathode. The ratio of anode



current, at an anode voltage sufficient to cause ionization, to the emission-limited current measured at a lower voltage is the gas amplification factor of a gas phototube. This factor ranges between 3 and 10 in commercial gas phototubes. See ELECTRICAL CONDUCTION IN GASES.

Gas amplification increases with anode voltage and with the intensity of incident radiation. Because of their nonlinear characteristics gas phototubes seldom are used in photometric applications. Gas amplification increases with anode voltage up to breakdown voltage, which is that voltage at which the ion current becomes self-sustaining. A self-sustained glow discharge in a gas phototube causes sputtering of the photocathode and rapidly impairs its sensitivity.

The response of a gas phototube to rapid changes in light intensity decreases with increasing modulation frequency, particularly at frequencies above 10–15 kilocycles. Factors which govern the speed of response of a gas phototube to modulated radiation input are ion transit time between anode and cathode, and delayed secondary emission produced at the photocathode by gas atoms which are excited to metastable energy states. The speed of response of a vacuum phototube is limited only by electron transit time across the interelectrode space.

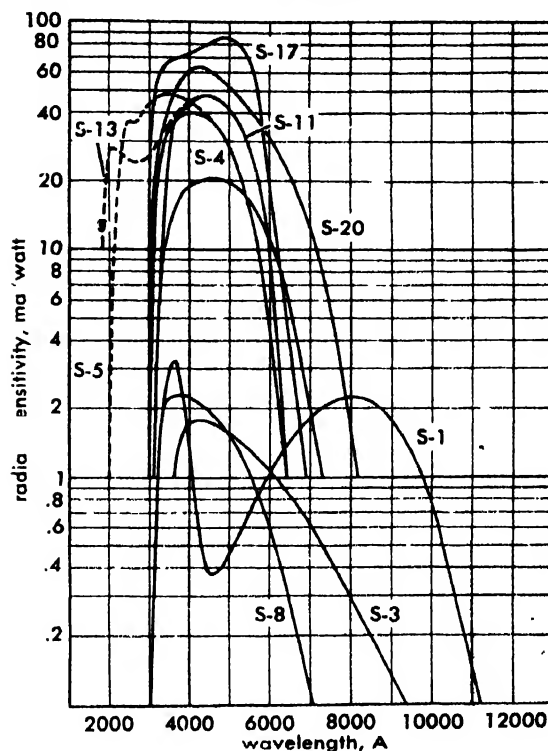
**Sensitivity to incident light.** The photocathode of a vacuum, gas, or multiplier phototube is selectively sensitive; that is, it emits electrons photoelectrically only when it is exposed to radiation having wavelengths in specific regions of the spectrum. Cathode radiant sensitivity is the photoelectric current emitted per unit of incident monochromatic radiant power. The spectral sensitivity characteristic of a phototube exhibits cathode radiant sensitivity as a function of the wavelength of radiation incident upon the window of the phototube.

Spectral sensitivity characteristics are shown in Fig. 2 and are designated by standard symbols S-1, S-3, etc. The sharp cutoff of sensitivity on the short-wavelength side of the curves is determined primarily by the transmission characteristic of the glass envelope or window of the phototube. The long-wavelength limit of radiant sensitivity is the threshold wavelength of the photocathode.

The sensitivity of a phototube is more easily measured as cathode luminous sensitivity, the photoelectric current emitted per unit of incident luminous flux from a specified source of light. It is expressed in microamperes per lumen. The source commonly used is a tungsten-filament lamp operated at a color temperature of 2870°K. Cathode-radiant and cathode-luminous sensitivities are measured with a collimated beam of radiation perpendicular to the window of the tube. Summarized in the table are radiant sensitivity at the wavelength of maximum response, luminous sensitivity, and other information relative to the spectral sensitivity characteristics represented in the illustration.

The quantum efficiency of a photocathode is the

number of electrons emitted per incident photon of a given wavelength. Because the energy per photon of wavelength  $\lambda$  is  $hc/\lambda$ , where  $h$  is Planck's constant and  $c$  is the velocity of light in vacuum, the quantum efficiency at a wavelength  $\lambda$  is simply the



Average spectral sensitivity characteristics of typical phototubes.

radiant sensitivity of the photocathode multiplied by the factor  $hc/\lambda$  in appropriate units. Typical quantum efficiencies are listed in the table.

**Photocathode material.** Photocathodes of practical importance contain one or more of the alkali metals lithium, sodium, potassium, rubidium, or cesium in complex combination with other metals or with oxides of certain metals. Because of their high reflectivity and conductivity, the pure alkali metals have lower quantum efficiencies than do the more complex photocathodes, which invariably are semiconductors.

Practical photocathodes may be classified broadly under two prototypes: the cesium oxide-silver cathode and the cesium antimonide cathode. The cesium oxide-silver cathode is obtained by permitting cesium to react with a thin layer of silver oxide. The resultant cathode layer is cesium oxide containing silver, possibly oxides more complex than  $Cs_2O$ , and a critical excess of cesium. Phototubes which contain this photocathode have the S-1 spectral sensitivity characteristic. The rubidium oxide-silver cathode is produced in a similar manner.

The cesium antimonide photocathode is obtained by exposing a thin layer of antimony to cesium vapor at elevated temperatures. The cathode surface is an intermetallic compound, cesium antimo-

## Average cathode characteristics

Spectral sensitivity characteristic*	Cathode material	Wavelength of maximum response, Å	Peak radiant sensitivity, ma/watt	Peak cathode quantum efficiency, %	Luminous sensitivity,† $\mu$ a/lumen	Remarks
S-1	Cs <sub>2</sub> O, Ag	8000	2.2	0.3	25	
S-3	Rb <sub>2</sub> O, Ag	4200	1.8	0.5	6.5	
S-4	Cs <sub>3</sub> Sb	4000	40	12.4	40	
S-5	Cs <sub>3</sub> Sb	3400	49	17.8	40	Ultraviolet transmitting window
S-8	Cs <sub>3</sub> Bi	3650	2.3	0.8	3	
S-10	Bi, Ag, O, Cs	4500	20.3	5.6	40	Semitransparent
S-11	Cs <sub>3</sub> Sb	4400	48	13.5	60	Semitransparent
S-13	Cs <sub>3</sub> Sb	4400	47	13.2	60	Semitransparent; ultraviolet transmitting window
S-17	Cs <sub>2</sub> Sb	4900	85	21.4	125	Semitransparent, on reflecting substrate
S-20	(NaKCs)Sb	4200	64	18.8	150	Semitransparent

\* These characteristics, shown in Fig. 2, refer to typical phototubes rather than to photocathodes.

† Light source is a tungsten-filament lamp operated at a color temperature of 2870°K.

nide, containing an excess of cesium. This cathode has maximum sensitivity in the blue and ultraviolet regions of the spectrum and a threshold wavelength at about 6500 angstroms (Å). The cesium-bismuth cathode is similar to the cesium antimonide cathode in composition and in method of preparation.

The bismuth-silver-oxygen-cesium cathode is formed by cesium activation of an oxidized layer of silver and bismuth. The S-10 characteristic associated with this type of photocathode is an effective combination of the blue response of the cesium-bismuth cathode and the red response of the cesium oxide-silver cathode. This broad characteristic, which extends over most of the visible spectrum, is a desirable feature for phototubes used in photometry and colorimetry.

The sodium-potassium-cesium-antimony, or tri-alkali, photocathode is produced by exposing a thin antimony layer to vapors of the alkali metals. This cathode has a higher peak radiant sensitivity than do any of the photocathodes mentioned above. Its spectral sensitivity characteristic extends from ultraviolet to infrared wavelengths. The tri-alkali photocathode is therefore an excellent panchromatic detector of visible and near-visible radiation.

**Photocathode construction.** A photocathode may be either an opaque layer of the emissive material on a metal electrode or a semitransparent layer on glass. A semitransparent layer is deposited directly on the window or envelope of the phototube. A portion of the layer overlaps a high-conductivity layer of aluminum or other metal which provides electrical contact to the photocathode. Radiation transmitted through the window is incident upon one side of the layer while electrons are emitted photoelectrically from the opposite, or vacuum, surface of the layer. This type of photocathode is commonly used in multiplier phototubes having S-1, S-10, S-11, S-13, or S-20 characteristics.

A semitransparent cathode may also be formed on an opaque, highly reflecting metal surface. Radiation transmitted through the cathode layer is reflected into and partially absorbed by a relatively thin layer from which photoelectrically excited electrons can readily escape. The high radiant sensitivity obtainable in this manner is illustrated by the S-17 spectral sensitivity characteristic.

**Dark current.** Dark current is the current measured at the terminals of a phototube when it is shielded from all radiation capable of causing photoelectric emission from its photocathode. Two common causes of dark current are electrical conductivity across or through the insulation supporting the electrodes or tube terminals, and thermionic emission from the photocathode. Electrical conductivity can be reduced to very small values by placing the cathode and anode terminals at opposite ends of the tube envelope, as illustrated in Fig. 1. Dark current due to thermionic emission is proportional to the area of the photocathode and increases almost exponentially with cathode temperature. At 25°C the thermionic emission from the cesium oxide-silver photocathode is of the order of  $10^{-11}$  amp/cm<sup>2</sup>, whereas that of the cesium antimonide photocathode is about  $1 \times 10^{-15}$  amp/cm<sup>2</sup>. At low light levels such that the photoelectric current and dark current are of similar magnitudes, discrimination between the two is readily achieved by modulating the input flux and measuring only the ac component of anode current.

**Types of service.** Phototubes are used in general in relay operation, detection of intensity-modulated radiation, and photometry. Response of a phototube to a change in light level causes a relay to close in safety and warning devices, and in counting or sorting equipment. Since the change in photocurrent may be considerably less than 1 microampere ( $\mu$ amp) in a vacuum phototube, a gas phototube is frequently used to trigger a thyratron in relay applications. Typical applications in-

involving the detection of intensity-modulated light are reproduction of sound-on-film and facsimile. Gas phototubes are commonly used in sound reproduction because of their inherent gas amplification and acceptable audio-frequency response characteristic. Vacuum phototubes are used in facsimile service for which high-frequency response is required.

Vacuum phototubes are used whenever linear response to radiant flux input is required as, for example, in photometry and in colorimetry. In photometric applications a spectral sensitivity characteristic similar to that of the human eye is frequently required and may be approximated by means of special optical filters used in combination with phototubes which are sensitive over the visible spectrum. In spectrophotometry two vacuum phototubes may be used, one having S-5 and the other S-1 response, in order to provide sensitivity at wavelengths from 2500 to 11,000 Å. In applications such as colorimetry and the measurement of density of color film, narrow-band filters are used with appropriate vacuum phototubes to establish test-stimulus values independently of the particular sensitivity characteristics of the tubes. These characteristics differ from tube to tube as a result of small uncontrollable variations in composition of the photocathode.

Vacuum and gas phototubes are capable of providing reliable service over thousands of hours of operation at moderate temperatures and radiation levels. Temperatures above 75–100°C and average cathode-current densities greater than about 30 amp/cm<sup>2</sup> cause a gradual decrease in sensitivity. The life of a phototube is limited largely by slow effusion of gaseous contaminants which cannot be baked out of the phototube after the photocathode is formed.

[J. L. WEAVER]

**Bibliography:** A. L. Hughes and L. A. DuBridge, *Photoelectric Phenomena*, 1932; A. H. Sommer, New photoemissive cathodes of high sensitivity, *Rev. of Scientific Instruments*, new ser., 26(7): 725–726, 1955; V. K. Zworykin and E. G. Ramberg, *Photoelectricity and Its Application*, 1949.

## Phototube, multiplier

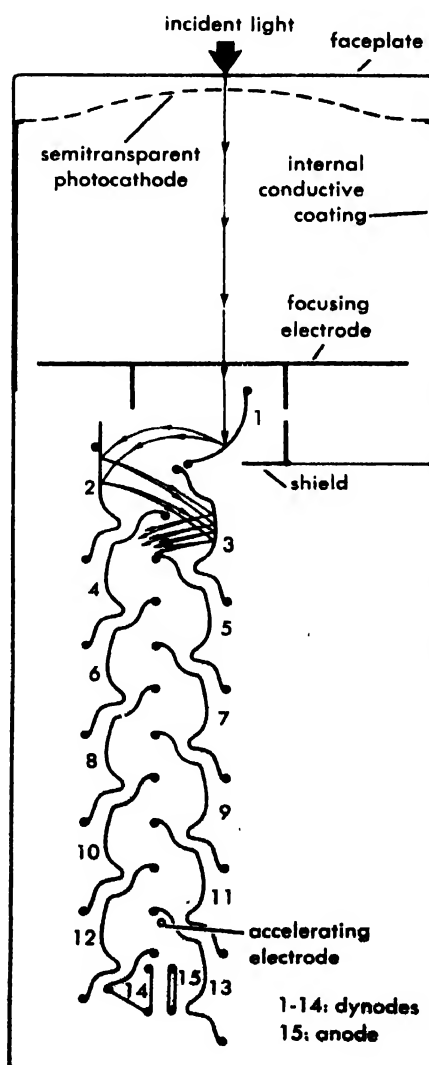
A phototube which contains one or more secondary-electron-emitting electrodes, or dynodes, between its photocathode and output electrode; also called photomultiplier tube. Electrons emitted by the photocathode initiate a cascade of secondary emission from dynode to dynode and ultimately to the output electrode, or anode. Because of the high current amplification thus obtained, a multiplier phototube is used in applications which require high photoelectric sensitivity and fast response to changes in the intensity of radiation input.

A multiplier phototube may be described by specifying its photocathode, the secondary-emission characteristic of its dynodes, and the type of dynode assembly or multiplier structure used in the tube. The photocathode may be any one of the

types discussed in connection with vacuum and gas phototubes. For a description of various photocathodes and their characteristics, see PHOTOTUBE.

**Operation.** In a multiplier phototube, electrons emitted from the photocathode are accelerated toward the first dynode, where they liberate secondary electrons. Similarly, these secondary electrons are directed toward the second dynode and cause emission of a larger number of secondary electrons. This process is repeated at all dynode stages of the multiplier structure which, typically, contains 6–14 stages. Each dynode is operated at a more positive potential than that of the dynode preceding it, the potential increasing by tens or hundreds of volts per dynode stage. In this manner, current-amplification factors of many millions are obtainable. The secondary emission from the last dynode is collected as anode current at the output electrode. When high amplification factors are involved, the output section of the multiplier must be well isolated from the input.

Depending on the method by which electrons are directed from dynode to dynode, multiplier struc-



Typical multiplier phototube construction.

tures may be classified as unfocused, electrostatically focused, and electromagnetically focused. In unfocused structures such as the grid, Venetian-blind, and box types, electrons are simply accelerated from dynode to dynode by means of grids. In electrostatically focused multipliers a portion of each dynode serves to shape the electric field between dynodes in such a manner that secondary emission from one dynode is focused upon the optimum area of the following dynode. Mutually perpendicular electric and magnetic fields provide similar focusing of secondary electrons in electromagnetically focused multipliers.

The transmission type of multiplier may be electrostatically or electromagnetically focused. In a transmission type of multiplier the dynodes are thin plane electrodes stacked in a parallel array. High-energy primary electrons incident on the metallic surface of each dynode are scattered in a thin metal film and cause secondary emission from a secondary-emissive layer on the opposite surface of the dynode. This type of multiplier requires potentials of a few kilovolts between successive dynodes in order to cause electrons to penetrate the metal film.

**Dynode coatings.** Of the many secondary-electron-emitting materials available, only a few are commonly used as dynode surfaces in multiplier phototubes. These materials have comparatively high secondary-electron emission coefficients at primary-electron energies of the order of 100 electron volts, and they can be used compatibly with certain photocathodes in the same envelope. Typical dynode surfaces are cesium antimonide on nickel or other metals, magnesium oxide on silver-magnesium alloy, and beryllium oxide on copper-beryllium alloy. The cesium-antimonide surface is formed by reaction of a thin antimony layer on the dynode with cesium vapor. Its secondary-emission coefficient is about 5 at primary-electron voltages between 100 and 120. The magnesium oxide and beryllium oxide surfaces are formed by surface oxidation of silver-magnesium and copper-beryllium alloys. These dynodes have a secondary-emission coefficient of approximately 3 for 100-volt primary electrons.

**Sensitivity.** The sensitivity of a multiplier phototube is the product of the photoelectric sensitivity of its photocathode and the current amplification of its multiplier structure. In the visible and near-visible regions of the spectrum the emission from typical photocathodes is of the order of 10–100 milliamperes per watt of incident radiation. The sensitivity of a high-gain multiplier phototube is accordingly many thousands of amperes per watt of radiation incident on its photocathode. At constant supply voltages the signal-output current of a multiplier phototube is directly proportional to radiant flux input over many orders of magnitude. However, average output currents greater than a few milliamperes cause a gradual decrease in secondary emission of the last few dynode stages, where current densities are highest.

Anode dark current of a multiplier phototube is the current observed at the anode in the absence of all radiation capable of causing photoelectric emission from its photocathode. Under normal conditions of tube temperature and supply voltage the dark current is predominantly amplified thermionic emission from the photocathode. This component of the anode dark current is proportional to the area of the photocathode. It can be reduced appreciably by cooling the cathode. The ultimate limit of detectability of a multiplier phototube is set by the shot noise associated with the amplified thermionic dark current. The rms noise current measured at the anode is proportional to the square root of the product of cathode emission and the bandwidth of the output circuit. The rms noise current per unit bandwidth resulting from thermionic emission in typical multiplier phototubes is equivalent to inputs as low as  $10^{-15}$ – $10^{-12}$  watt of modulated radiation incident on the photocathode at normal ambient temperatures.

Multiplier phototubes are used in the detection of very low levels of visible and near-visible radiation. Typical uses include astronomical star tracking, low-level photometry, spectrometry, flying spot television pickup, and scintillation counting. In scintillation counting a multiplier phototube with a semitransparent photocathode is used in conjunction with a scintillating material to produce output current pulses which are proportional to the energy of  $\gamma$ -quanta or nuclear particles capable of exciting the scintillator. In these applications a multiplier phototube has a useful operating life of thousands of hours. However, small changes in secondary emission at each dynode, when compounded, produce appreciable changes in sensitivity during continuous operation. [J. L. WEAVER]

**Bibliography:** G. A. Morton, Photomultipliers for scintillation counting, *RCA Rev.*, 10(4):525–551, 1949; E. J. Sternglass, High speed electron multiplication by transmission of secondary electron emission, *Rev. Sci. Instr.*, 26(12):1202, 1955; V. K. Zworykin and E. G. Ramberg, *Photoelectricity and Its Application*, 1949.

## Photovoltaic cell

A device that detects or measures electromagnetic radiation by generating a potential at a junction (barrier layer) between two types of material upon absorption of radiant energy. Typical junctions for photovoltaic cells are silicon-silicon-boride, selenium-iron, copper oxide-copper.

The detection or measurement is performed by connecting the cell directly to a galvanometer, whose reading is a function of the intensity of radiation falling on the cell.

Photovoltaic cells are used as exposure meters in photography and are used in automation to energize sensitive relays. The solar battery, a type of photovoltaic cell, may be used as a source of electricity for portable radios and telephone relays.

A photovoltaic cell has practically no dark current. It is generally not adapted to be used with

amplifiers. See PHOTOELECTRIC DEVICES; PHOTOVOLTAIC EFFECT; SOLAR BATTERY. [J. J. ROBILLARD]

**Bibliography:** Stanford Research Institute, *Proceedings World Symposium on Applied Solar Energy*, 1956.

### Photovoltaic effect

A term most commonly used to mean the production of a voltage in a nonhomogeneous semiconductor, such as silicon, by the absorption of light or other electromagnetic radiation. In its simplest form, the photovoltaic effect occurs in the common photovoltaic cell, used, for example, in solar batteries and exposure meters. The photovoltaic cell consists of an *n-p* junction between two different semiconductors, an *n*-type material in which conduction is due to electrons, and a *p*-type material in which conduction is due to positive holes. When light is absorbed near such a junction, new mobile electrons and holes are released, as in photoconduction. An additional feature of a photovoltaic cell, however, is that there is an electric field in the junction region between the two semiconductor types. The released charge moves in this field. This current flows in an external circuit without the need for a battery as required in photoconduction. If the external circuit is broken, an "open-circuit photovoltage" appears at the break.

In certain rather complex electrolytic systems, illumination of the electrodes may give rise to a voltage classed as photovoltaic. See EXPOSURE METER; PHOTOCONDUCTIVITY; PHOTOVOLTAIC CELL; SEMICONDUCTOR; SOLAR BATTERY. [L. APKER]

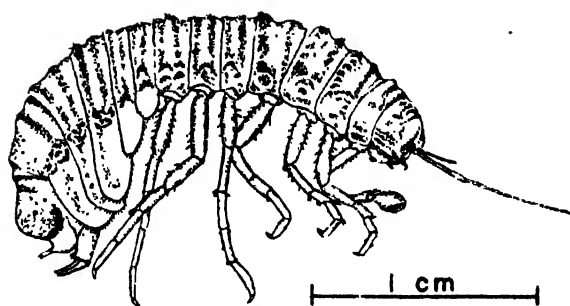
**Bibliography:** A. Van der Ziel, *Solid State Physical Electronics*, 1957.

### Phreatoicoidea

A suborder of the Isopoda and class Crustacea. The body is sybicylindrical, appearing laterally compressed, mainly because of the downward development of the pleura of the pleon. The first and occasionally the second thoracic segment is fused with the head. Antennules are shorter than the antennae. The eyes may be large, small, or absent. Mouthparts are primitive. The first four pairs of pereopods are directed forward, while the posterior three pairs are directed backward. The first pair is subchelate. The pleon has six distinct segments with the last being fused with the telson but marked from it by a suture. Pleopods are broad, foliaceous and branchial in function while the uropods are biramous and lateral. The suborder is divided into two families, the Amphisopidae, having both mandibles with a lacinia mobilis, and the Phreatoicidae, in which only the left mandible retains a lacinia mobilis.

The suborder is an ancient one and includes a fossil *Protamphisopus wianamattensis* (Chilton) from the Triassic beds of New South Wales. Three extant species are recorded from Australia, Tasmania, New Zealand, and South Africa and one that is subterranean from India.

Most species occur in fresh water. Several are



*Onchotelson brevicaudatus* (Smith) adult male.

blind, subterranean forms and one occurs in hot water from deep artesian bores. A few are semiterrestrial, burrowing forms. [E. M. SHEPPARD]

**Bibliography:** G. Nicholls, *Papers and Proc. Roy. Soc. Tasmania*, 1943:1-145, 1944:1-157; E. Sheppard, *Proc. Zool. Soc. London*, 1927:1, 81-124.

### Phrynophiurida

An order of Ophiuroidea in which the vertebrae usually articulate by means of hourglass-shaped surfaces and the arms are able to coil upward or downward in the vertical plane. There is usually a leathery integument, in which calcareous granules or platelets are imbedded. Most species are found in deep water, and often the arms are tightly coiled about the branches of black corals, upon which Phrynophiurida feed. Of the three families, the Gorgonocephalidae often have branched arms, the Asteronychiidae have a large disk and slender arms, and the Asteroschematidae have a small disk and stout arms.

The foregoing families share a number of characteristics enabling their union in one suborder, Euryalina. One remaining family, the Ophiomyxidae, differs in having a soft, unprotected integument, like that of *Ophiocanops*, but lacks the peculiar features of the gut and gonads in oegophiurids. For reasons too specialized to discuss here, it appears best to associate the Ophiomyxidae with the Euryalina, in the Phrynophiurida, placing the family in a distinct suborder, the Ophiomyxina. See OEGOPHIURIDA; OPHIUROIDEA. [H. B. FELL]

**Bibliography:** H. B. Fell, Evidence for the validity of Matsumoto's classification of the Ophiuroidea, *Publ. Seto Marine Biol. Lab.*, 10:145-152, 1962.

### Phthalic acid

One of the three benzenedicarboxylic acids of formula  $C_6H_4(COOH)_2$ . *o*-Phthalic (or simply, phthalic) acid, melting point  $191^\circ C$  (sealed tube) is the 1,2 isomer; isophthalic (or *m*-phthalic) acid, melting point  $347-348^\circ C$ , is the 1,3 isomer; terephthalic acid, melting point  $425^\circ C$  (sealed tube) is the 1,4 isomer.

The acids are prepared by permanganate or chromic acid oxidation of appropriate xylenes, or by partial chlorination of the xylene, followed by basic hydrolysis and finally oxidation. *o*-Phthalic

acid, commercially the most important of the three, is manufactured mainly by vapor-phase oxidation of naphthalene over a vanadium pentoxide catalyst at about 480°C, whereby the product, phthalic anhydride, sublimes from the reaction zone in a state of high purity. See OXIDATION PROCESS.

Phthalic acids are used in chemical analysis, in preparation of esters (methyl, ethyl, and butyl phthalates), and in preparation of phthaloyl chlorides,  $C_6H_4(COCl)_2$ , from the acid and phosphorus pentachloride. Phthalic anhydride can be decarboxylated to benzoic acid, and can be used in the synthesis of indigo and derivatives of anthraquinone. Phthalic anhydride reacts at elevated temperature with polyalcohols (ethylene glycol or glycerol) to form polyesters which are used as plastics. With ammonia, phthalic anhydride gives phthalimide. Terephthalic acid, heated with ethylene glycol, gives polyesters used as synthetic fibers. See ACID ANHYDRIDE; CARBOXYLIC ACID; PHTHALIMIDE; POLYESTER RESINS; XYLENE. [E.B.R.]

## Phthalimide

The imide of *o*-phthalic acid, also called 1,3-isobenzodione. The melting point of phthalimide is 238°C, and it is only slightly soluble in water. It is a weak acid,  $K_a = 5 \times 10^{-9}$ . The substance is prepared commercially by the reaction of molten phthalic anhydride and ammonia; in the laboratory, phthalic anhydride and either ammonium hydroxide or ammonium carbonate are used.

In the form of its sodium or potassium salt, it is widely used in the synthesis of both primary amines and amino acids. It is also combined with the malonic ester in the synthesis of complex amino acids.

Phthalimide is used as starting material in the synthesis of methyl anthranilate, the active principle in jasmine and orange oils. See AMIDE, ACID; AMINE; PHTHALIC ACID. [E.B.R.]

## Phycomycetes

A class of fungi of the subdivision (or, in the opinion of some mycologists, division) Eumycetes or Eumycophyta. The Phycomycetes are the most primitive of the true fungi. Many species live in the water and are superficially similar to certain green algae. In addition to the aquatic Phycomycetes there are other species that live in the soil. Modern studies indicate that the Phycomycetes evolved independently, along several lines, from colorless flagellate ancestors. Some phycomycetous species cause disease in crops, man, and animals; some parasitize insects and fish; some cause food damage in the home and various fruit and vegetable rots in transit and storage; other species are used in industrial fermentation processes.

As in all Eumycophyta, the growing, vegetative thallus is enclosed in a definite cellular wall, so that nutrient materials must enter the organism in solution. Two morphological features distinguish the Phycomycetes from other classes of fungi. (1) The actively growing portions of the plant

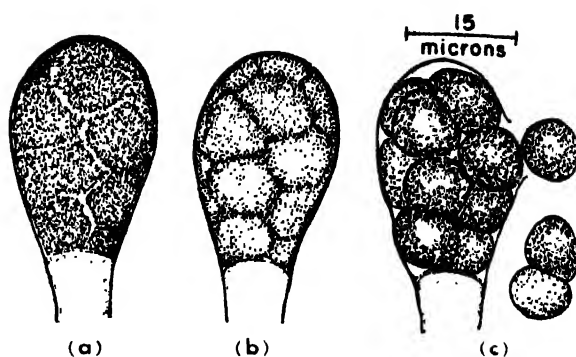


Fig. 1. Progressive cleavage in a sporangium of the water mold *Thraustotheca*. (a) First cleavage planes are just apparent. (b) Cleavage planes nearly completed. (c) Fully formed sporangiospores being released from sporangial wall. (From Fitzpatrick, 1930, after Weston)

body lack regularly spaced septa or cross walls. Septa are generally formed only where reproductive structures arise or in older relatively inactive regions of the mycelium. (2) The fundamental, asexual reproductive unit is the sporangiospore produced in the sporangium by a distinctive process termed progressive cleavage (Fig. 1). Starting with a multinucleate or coenocytic mass of protoplasm contained in the young sporangium, a pattern of cross walls is progressively formed until the entire protoplast is divided into units, each of which ultimately matures into a spore. Although each unit usually has a single nucleus, binucleate or multinucleate spores are formed exceptionally in certain species and regularly in others.

**Morphology and reproduction.** The phycomycetous thallus or plant body ranges from a single globular cell without branches of any sort (Fig. 2a) in the simplest of the Chytridiales, to globose, elongate, or branched forms with basal rhizoids (Figs. 2b and 4), and finally to a typical, extensively branched and often cottony mycelium in the Peronosporales and Mucorales.

Reproductive processes are also highly variable. Sexuality is widespread and involves the production of one or more gametes or sex cells in each gametangium just as spores are formed in a sporangium. Syngamy, which is the fusion of two gametes, takes place through a number of different mechanisms. One of these involves the release and subsequent fusion of motile gametes. If these are morphologically indistinguishable, they are referred to as plus (+) and minus (−) types; if they are of different size, the smaller one is indicated as the male (♂), the larger as the female (♀) gamete (Fig. 3a,b). Other mechanisms involve the fertilization of a large nonmotile egg by a small free-swimming sperm cell (Fig. 3c), the passage of male nuclei to the eggs through fertilization tubes (Fig. 3d), or the direct fusion of like or unlike gametangia (Fig. 5b). Syngamy often occurs between sex cells from a single hermaphroditic thallus but in many other instances fusion will occur



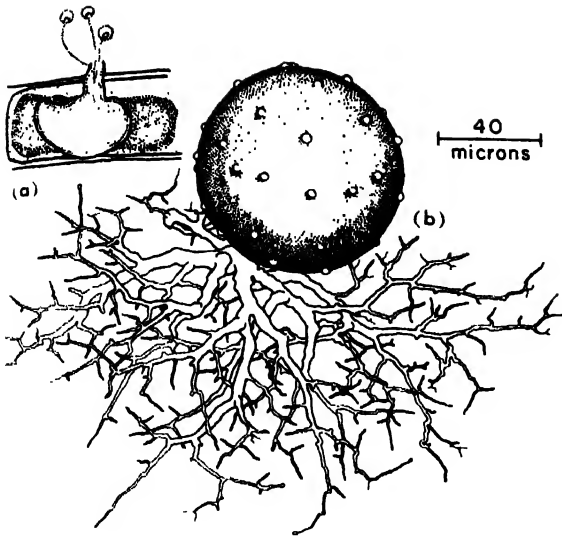


Fig. 2. (a) One-celled, saclike thallus of *Olpidium* within an algal cell; entire thallus has been converted into a sporangium and has emptied its zoospores (from Sparrow, 1943). (b) The thallus of *Rhizophydium* is made up of a sporangial sac with a basal tuft of anchoring and absorptive rhizoids (drawing by F. V. Ranzoni).

only between the products of two different thalli. The former situation is known as homothallism, the latter is known as heterothallism. See SYNGAMY.

Meiosis involves halving of the chromosome number and usually occurs in the zygote; thus, the Phycomycetes are, like most other fungi, generally haploid. Postponed meiosis leading to true diploidy and an alternation of gametophytic and sporophytic generations occurs in the genus *Allomyces* of the Blastocladales (Fig. 4). See MEIOSIS; METAGENESIS.

The asexually formed sporangiospores are produced in great numbers and serve for rapid multiplication and spread. They are usually flagellated and motile in the aquatic Phycomycetes and are called zoospores. The number and arrangement of the flagella have provided a sound basis for grouping the orders concerned (see table). Electron microscopy shows that flagella of the Phycomycetes have the 11-strand structure found throughout the plant and animal kingdoms.

In the Peronosporales, paralleling adaptation to a terrestrial life, there has been a progressive trend toward deciduous sporangia which are forcibly discharged from specialized sporangiophores. Such sporangia, instead of cleaving into zoospores, frequently germinate directly by a hyphal tube. They are then spoken of as conidia but their basic sporangial character is clear. The sporangia of the Entomophthorales are also forcibly discharged as a rule and function as conidia.

The septum that separates the sporangium from the sporangiophore in the Mucorales is frequently arched upward forming a dome-shaped columella (Fig. 5a). Numerous nonmotile sporangiospores are

released and passively distributed by wind and rain. Whole sporangia are shot several feet by a remarkable explosive mechanism in *Pilobolus*. Deciduous conidia are produced in many genera of the Mucorales and their evolution from sporangia can be traced through loss of the columella and progressive reduction in size until only one spore is contained within the sporangial wall.

A resistant stage of some sort occurs in most Phycomycetes and serves to tide them over periods of adversity. Usually it is the zygote, provided with a thickened, pigmented wall and a rich food supply, that fulfills this function and remains dormant, sometimes for years. The oospores (Fig. 3d) of the Saprolegniales and Peronosporales and the zygospores (Fig. 5b) of the Mucorales are good examples. In the Blastocladales and possibly a few other groups, resistant sporangia are formed on a sporophytic thallus. In many genera of Phycomycetes dense masses of protoplasm also become walled off to form resistant gemmae or chlamydospores.

**Ecology and physiology.** The Phycomycetes occur all over the world in soils, in fresh waters, and

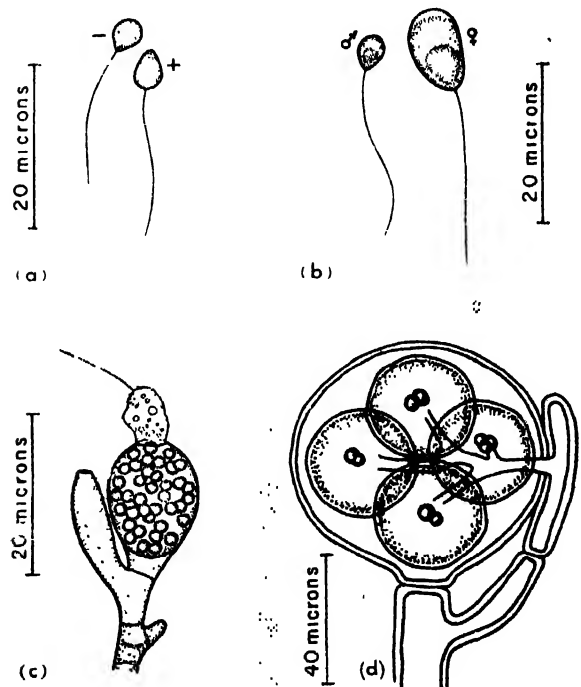


Fig. 3. Types of gametes in the aquatic Phycomycetes. (a) Equal, motile gametes of *Olpidium* (after Kusano, redrawn from H. Kniep, *Die Sexualität der niederen Pflanzen*, 1928). (b) Unequal motile gametes of *Allomyces*. (c) Motile sperm just entering the large, nonmotile egg of *Monoblepharella* (from Sparrow, 1943). (d) Gametes of *Saprolegnia*; the male nuclei are carried to the nonmotile eggs through fertilization tubes and the fertilized eggs then develop into resistant oospores (semidiagrammatic, after Clausen from H. Kniep, *Die Sexualität der niederen Pflanzen*, 1928).



**Tabular summary of the Phycomycetes**

Order	Habitat	Thallus	Spores	Representative genera
Chytridiales	Mainly aquatic	Saclike to rhizoidal	Zoospores with one posterior flagellum	<i>Olpidium</i> , <i>Rhizophydium</i> , <i>Synchytrium</i>
Blastocladales	Aquatic	Basal rhizoids and terminal hyphae	Zoospores with one posterior flagellum	<i>Allomyces</i> , <i>Blastocladiella</i>
Monoblepharidales	Aquatic	Mostly hyphal	Zoospores with one posterior flagellum	<i>Monoblepharella</i> , <i>Monoblepharis</i>
Hypochytriales	Aquatic	Saclike to limited hyphal	Zoospores with one anterior flagellum	<i>Rhizidiomyces</i> , <i>Hypochytrium</i>
Saprolegniales	Aquatic	Mostly hyphal	Zoospores with two flagella	<i>Saprolegnia</i> , <i>Achlya</i> , <i>Aphanomyces</i>
Leptomitales	Aquatic	Hyphal, or basal rhizoids and terminal hyphae	Zoospores with two flagella	<i>Leptomitul</i> , <i>Sapromyces</i>
Lagenidiales	Aquatic	Saclike to limited hyphal	Zoospores with two flagella	<i>Olpidiopsis</i> , <i>Lagenidium</i>
Peronosporales	Aquatic to terrestrial	Hyphal	Zoospores with two flagella, or conidia	<i>Pythium</i> , <i>Phytophthora</i> , <i>Peronospora</i>
Entomophthorales	Mainly terrestrial	Hyphal	Nonmotile sporangiospores, or conidia	<i>Entomophthora</i> , <i>Basidiobolus</i>
Mucorales	Terrestrial	Hyphal	Nonmotile sporangiospores, or conidia	<i>Mucor</i> , <i>Rhizopus</i> , <i>Pilobolus</i>

in the oceans. They grow saprophytically on every sort of natural organic material and parasitically on every major group of plants and many kinds of animals. Some are generalized scavengers while others prefer particular substrata like submerged fruits or the dung of herbivores, and still others concentrate upon cellulose, chitin, or hair. They range from weak facultative forms to the most highly obligate parasites. Among the serious crop diseases they cause are late blight of potatoes, seedling and storage rots, and the iniquitous downy mildews of grapes, tobacco, lettuce, and many other important economic plants. As insect parasites the Entomophthorales play a role in natural control of insect infestations. The fish mold, *Saprolegnia*, is sometimes a pest in aquaria and fish hatcheries, and certain species of Mucorales have been implicated in mycoses of higher animals and man. See INSECT PATHOLOGY; MYCOLOGY, MEDICAL; PLANT DISEASE; SOIL MICROBIOLOGY.

Except for the obligately parasitic Peronosporales, most of which remain to be cultured, many genera in all the major groups have been studied in pure culture and much is known about their nutrition and metabolism. Like most fungi the Phycomycetes are generally obligate aerobes. However, Louis Pasteur showed that *Mucor* can cause alcoholic fermentation, and more recent evidence indicates that a number of the aquatic genera are actually facultative anaerobes, being able to grow in the presence or absence of oxygen. Lactic acid is apparently one of the common products of metabolism along with succinic acid, acetaldehyde, and various other compounds. Because of their strong starch-splitting capacity, species of *Rhizopus* have

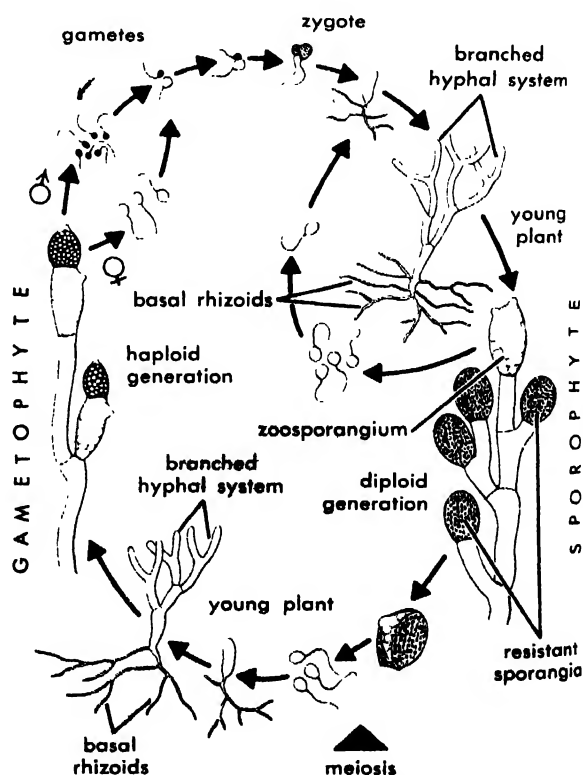


Fig. 4. Life cycle and reproductive behavior of the water mold *Allomyces*. Diploid spore-bearing plants, the sporophytes, alternate with the haploid gamete-bearing plants, the gametophytes. The gametophytes are hermaphroditic and self-fertile. (Semidiagrammatic, from A. T. Brice, *Syngamy and Alternation of Generations in Allomyces*, a phase in 1953)

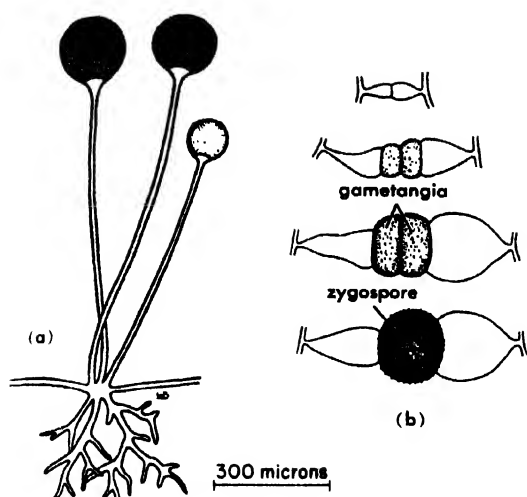


Fig. 5. Reproduction in the Mucorales. (a) Three sporangial stalks and sporangia of the bread mold *Rhizopus*; the columella is shown within the spore mass of the two larger, mature sporangia (from H. J. Fuller and O. Tippo, *College Botany*, rev. ed., Holt, 1954). (b) Stages in the sexual reproduction of *Rhizopus*; direct fusion of gametangia results in the formation of a heavy-walled zygospore. (From M. C. Coulter, *The Story of the Plant Kingdom*, Univ. Chicago Press, 1935)

been used industrially to convert starchy materials to sugar in the amylo process for alcohol production. See ETHYL ALCOHOL.

Whereas some phycmycetous species need only an organic source of carbon and energy, many have more complex growth requirements. Thiamin is a widespread requirement throughout the group and *Phycomyces* has long been used as a bioassay organism for this vitamin. Other demonstrated nutrient requirements, mostly among the aquatic genera, are p-aminobenzoic acid, biotin, nicotinamide, methionine, inositol, and an organic iron-complex the nature of which is still unclear. While few definitive chemical studies have been made, it is suggested that the cell walls of the *Phycomyces* are probably chitinous in most genera, cellulosic in some, and possibly a combination of the two in others.

The *Phycomyces* have provided material for many fundamental biological studies. Noteworthy among these are recent investigations relating to hormone-controlled sexuality in *Achlya*; hybridity, differentiation, and reproductive behavior in several genera of the Blastocladales; and light-induced growth responses in various Mucorales. [R.E.]

**Bibliography:** E. A. Bessey, *Morphology and Taxonomy of Fungi*, 1950; H. M. Fitzpatrick, *The Lower Fungi, Phycomyces*, 1930; F. K. Sparrow, Jr., *Aquatic Phycomyces*, 2d ed., 1960.

## Phylactolaemata

A class of the phylum Bryozoa. Though numbering fewer than 50 species, the class is cosmopolitan, occurring in fresh waters of every continent and

in many climates. The colonies, with some exceptions, die at the onset of winter. New colonies arise by germination of dormant, seedlike bodies, called statoblasts, which have remained viable over winter or dry seasons. Statoblasts of each family are distinctive (Fig. 1).

The most commonly encountered family, Plumatellidae, has sessoblasts and spineless floatoblasts. Its mosslike colonies are firm, yellowish to brown, chitinous or encrusted, with tubular branching zooids topped by tentacled polypides.

Cristatellidae and Lophopodidae form soft, colorless, gelatinous, baglike colonies having limited power of locomotion—some creep along the substratum. They have spinoblasts but no sessoblasts.

Fredericellidae resemble the plumatellids in colony appearance but have only sessoblasts or ptioblasts.

Most phylactolaematous lophophores are horse-shoe-shaped (Fig. 2) and the tentacles number 16–106. A flexible, muscular, hollow ciliated fold, the epistome, overhangs the mouth on the neural

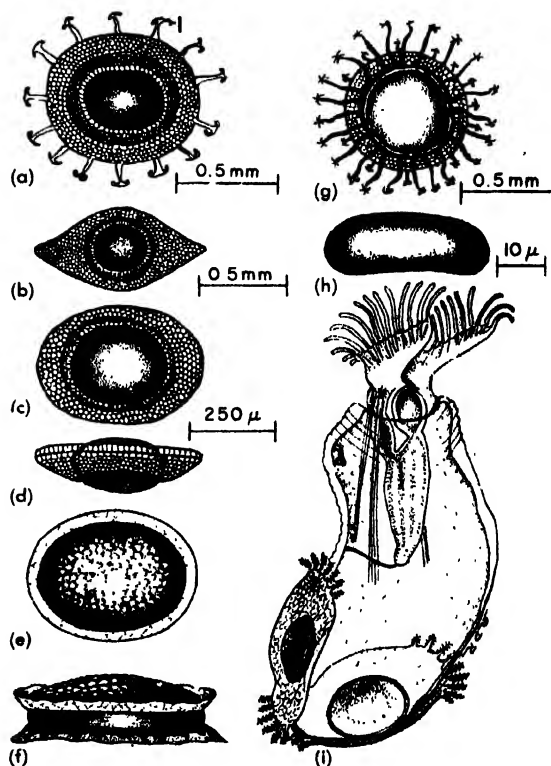


Fig. 1. Statoblasts of Phylactolaemata. (a) *Pectinatella magnifica* spinoblast. (b) *Lophopus crystallinus* spinoblast. (c) *Plumatella repens* floatoblast. (d) *Hyalinella punctata* floatoblast, edge view; face view would be similar to the *Plumatella* floatoblast in (c). (e) *Stolella indica* sessoblast, face view; annulum vestigial, not a true float. (f) *Stolella indica* sessoblast, edge view. (g) *Cristatella mucedo* spineblast. (h) *Fredericella sultana* sessoblast or ptioblast. (i) Germinated *Lophopodella carteri* statoblast (spinoblast), the growth of the first new zooid pushing apart its two valves.

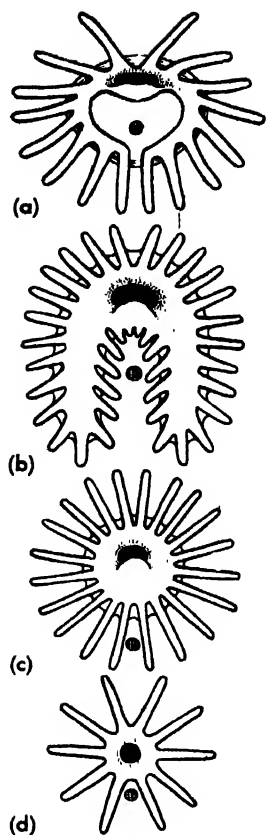


Fig. 2. Bryozoan tentacular crown, top view, diagrammatic. (a) *Loxosoma*, an entoproct. (b) *Plumatella*, a phylactolaematous ectoproct. (c) *Fredericella*, a phylactolaematous ectoproct. (d) *Bowerbankia*, a gymno-laematous ectoproct.

side. The gut is V-shaped and attached to the body wall by a funiculus, a thin ribbon of tissue from which statoblasts and sperms develop. The ovary is higher up on the body wall. See BRYOZOA; LOPHOPHORE; STATOBLASTS. [M.D.RO.]

## Phyllite

A large group of regional metamorphic rocks derived from argillaceous sediments and recrystallized in the greenschist facies (low degree of metamorphism), essentially composed of white mica and quartz. Phyllites are fine-grained, strongly schistose rocks, and the schistosity surfaces exhibit a glittery sheen given off by mica. They are widely distributed and easily recognized. In the past, geologists called them the lustrous schists of the mountain ranges (the *schistes lustrés* of the French and Swiss Alps). With very low metamorphism the phyllites pass into slates, and with increasing metamorphism they pass into mica schists. See METAMORPHIC ROCKS; SCHIST; SLATE.

The simple mineral composition of the ordinary phyllite (quartz and muscovite) is explained by the rule of the paucity of mineral phases. With increasing content of iron-magnesia, new minerals like chlorite, almandine-spessartite garnet, or chlo-

ritoid may develop. Among the garnetiferous phyllites may be mentioned the whetstones of the Ardennes. Chloritoid minerals are present in the ottrelite schists of the Alps. Paragonite, the sodium-mica, is commonly present in phyllites.

Phyllite has a wide distribution in the crystalline mountain ranges of the world: the Highlands of Scotland, mountains of Norway, Harz Mountains and mountains of Saxony, the Alps, Appalachians, and Great Lakes district of the United States. The schistosity of phyllites is sometimes flat but usually is crumpled and so imperfect as to render the rock unsuitable for roofing material. [T.F.W.B.]

## Phyllocarida

This term has two rather different meanings in crustacean literature. In the original wider sense, the division Phyllocarida includes both recent and fossil representatives of, and is thus synonymous with, the Malacostracan series Leptostraca. In the narrower sense, adopted by many palaeontologists, the fossil forms are placed in the order Phyllocarida, the order Nebaliacea being restricted to the recent ones. See LEPTOSTRACA; NEBALIACEA. [I.G.O.]

## Phyllonite

A metamorphic rock—the name is a combination of phyllite and mylonite; the phyllonites occupy a position between the rock types representing the component parts of the name. There are two distinct stages of their development. In the first stage the original rock is granulated by extreme deformation and pulverized to a mylonite. In the second stage, but frequently overlapping the first stage in time, new minerals recrystallize and grow (this is called crystalloblastesis). B. Sander originally introduced the name for phyllitelike rocks which had suffered a deformation after recrystallization regardless of whether they were derived from argillites, like the phyllites, or from orthocataclastites. See METAMORPHIC ROCKS; MYLONITE; PHYLLITE. [T.F.W.B.]

## Phymosomatoida

An order of Echinacea with a stirodont lantern and diademoid ambulacral plates; each interambulacral plate typically carries more than one primary tubercle (see ECHINOIDEA). There are three families. The Pseudodiademmatidae, of the Jurassic and Cretaceous had perforate crenulate tubercles. The Phymosomatidae, with imperforate crenulate tubercles, arose in the Jurassic. One surviving genus, *Glyptocidaris*, occurs off Japan. The Stomopneustidae, with imperforate noncrenulate tubercles, also range from the Jurassic, and the sole surviving genus, *Stomopneustes*, is a common Indo-Pacific littoral form. See ECHINACEA. [H.B.F.]

## Physical chemistry

Physical chemistry lies between physics and chemistry. It provides the theoretical basis of chemistry, and explains and predicts phenomena in all

branches of chemistry, including inorganic, organic, and analytical chemistry, and biochemistry. It is the foundation of chemical engineering. When the emphasis is more exclusively on physics, the subject is sometimes called chemical physics.

Physical chemistry covers a wide variety of subjects and it uses many different techniques. It includes studies of chemical equilibrium, reaction rates, solutions, molecular weights, molecular structure, and the properties of gases, liquids, crystals, and colloids. It involves the influence of temperature, pressure, electricity, light, concentration, and other physical factors on chemical systems.

Physical chemistry embraces laboratory measurements of chemical systems, mathematical descriptions, and theoretical interpretations. There is a constant evolution and improvement of apparatus to make possible greater accuracy in laboratory measurements.

There are three different approaches to the study and use of physical chemistry: thermodynamics, which is based on energy and which depends on the statistical behavior of a large number of molecules under equilibrium conditions; kinetics, which involves chemical changes with time; and molecular structure, which correlates the arrangement of atoms and electrons with chemical and physical properties and with chemical changes.

In describing and predicting chemical behavior, physical chemistry makes extensive use of graphs and mathematical formulas. Many of these formulas are based on calculus and differential equations, and on quantum and statistical mechanics. The various branches of physical chemistry employ special techniques, but they all involve the general principles just mentioned.

**Energy relationships.** Chemical thermodynamics is concerned with the relation between energy and chemical and physical change. The energy changes are easily determined, and they give a quantitative measure of the changes involved. Heat, work, pressure, temperature, and volume are determined directly, and they, in turn, lead to calculations of the thermodynamic quantities: enthalpy  $H$ , internal energy  $E$ , entropy  $S$ , free energy  $F$  or  $G$ , and work content  $A$ . One of the most useful relations in chemical thermodynamics is  $\Delta F^\circ = -RT \ln K$ , which permits a calculation of the equilibrium constant  $K$  and the determination of how far a chemical reaction will go. In this formula,  $\Delta F^\circ$  is the change in standard free energy of the reactants and products,  $R$  is the gas constant,  $\ln$  is the logarithm to the base  $e$ , and  $T$  is the absolute temperature. The free energy change can be calculated from the change in enthalpy (the heat of reaction at constant pressure) and the change in entropy, both of which are obtainable by calorimetric measurements. Quantitative calculations are made with the help of statistical mechanics. See THERMODYNAMICS (CHEMICAL).

Thermochemistry involves the heat evolved or absorbed by chemical reactions and by changes in

temperature. The heat changes are measured in calorimeters. The heats of formation of chemical compounds from their elements under standard conditions are recorded in tables. By subtracting the heats of formation of the reacting materials from those of the products, it is possible to calculate the heat of chemical reactions. The heats of reaction at different temperatures are calculated from the heat capacities of the reactants and products. The heats of reaction are useful, not only for making physical chemical calculations, but for practical purposes such as the evaluation of the heating values of fuels.

**Gases, solids, and liquids.** The physical chemical study of gases includes equations of state which give the relation between absolute temperature  $T$ , pressure  $p$ , and volume  $v$ . The simplest equation of state is

$$pv = nRT$$

where  $n$  is the number of moles, and  $R$  is the gas constant expressed in proper units—usually liter-atm/(deg) (mole). More elaborate formulas are needed for greater precision or for high pressures. The distribution of velocities at a given temperature among a large number of molecules is given by the Maxwell-Boltzmann distribution law. The density of a gas, obtained from measurements of weight per unit volume at definite temperatures and pressures, permits a calculation of the molecular weight of the gas. See GAS.

Crystals are made up of atoms or ions arranged in an orderly, repetitive manner in a solid. Much can be learned about the arrangement of the atoms within the crystal and the distance between them by measuring the diffraction of a beam of x-rays passing through the crystal. Important predictions concerning the properties of the crystals can be made from a knowledge of the crystal structure. See CRYSTAL STRUCTURE; SOLID-STATE CHEMISTRY.

The study of liquids is more complicated than that of gases or crystals. The properties of liquids studied in physical chemistry include freezing and vaporization, vapor pressures at different temperatures, density, surface tension, viscosity, and dielectric constant. See LIQUID.

**Solutions.** Solutions of solids in liquids and of liquids in liquids constitute an important branch of physical chemistry. A knowledge of vapor pressures, of liquid mixtures and of fractional distillations is necessary for the effective separation of liquids. Solubilities of solids in liquids at various temperatures are important. The influence of the dissolved substance in lowering the freezing point, raising the boiling point, and lowering the vapor pressure of the solvent are related quantitatively to the molecular weight and the concentration of the dissolved material through thermodynamic formulas. The creation of an osmotic pressure is another physical-chemical phenomenon of solutions. See SOLUTION.

**Equilibria.** Chemical equilibria are important for calculating how far chemical reactions will go and

what percentage yield of new products can be obtained at different temperatures and pressures. The equilibrium constants are determined experimentally by measuring the concentrations of the reactants and products at equilibrium, or they are calculated with the help of thermodynamics. The equilibria are determined in gases or in liquid solutions.

If the system contains gases alone, or only one liquid solution, so that only one phase is present, the equilibrium is said to be homogeneous. If the system contains more than one phase (gas, solid, or liquid), the equilibrium is said to be heterogeneous. Phase diagrams describe the influence of concentration and temperature on the number of solid and liquid phases. The phase rule is useful in correlating the number of phases and the number of independent variables. See EQUILIBRIUM, CHEMICAL.

**Kinetics.** Chemical kinetics is that branch of physical chemistry which is concerned with the prediction of reaction rates and with understanding the mechanism of chemical reactions. In some reactions, the rate depends directly on the concentration of the reacting material and in others on the square of the concentration. Most reactions are quite complex, with several different reactions going on together so that the relation between concentration and rate is difficult to determine. The influence of temperature on reaction rate is important also. It is large, most chemical reactions at room temperature doubling in rate for each 10°C rise in temperature. See KINETICS (CHEMICAL).

**Electrochemistry.** The electrical conductance of solutions provides important information in physical chemistry. Most salts and inorganic acids and bases dissociate into electrically charged ions when dissolved in water. When charged electrodes are placed in such a solution, the positive ions migrate to the negative electrode and the negative ions to the positive electrode. The conductivity of the solution depends both on the number of ions and on their velocity. When the ions reach the electrode, they may undergo a chemical change to gain electrons at the cathode or to lose them at the anode. There is an exact relation between the quantity of electricity and the extent of the chemical change. According to Faraday's law, 1 gram-equivalent weight of material is electrolyzed for each 96,500 coulombs of electricity.

Equilibrium relations involving ions give important information on hydrolysis, solubility, electrolytic dissociation, and the formation of complex ions.

The electromotive force or voltage of an electrochemical cell is an important quantity in physical chemistry. When suitable electrodes are surrounded by oxidizing or reducing ions and arranged in pairs, definite voltages are generated which can be used for calculating the free energy of the chemical reaction involved and the corresponding equilibrium constant. The potential between a single electrode and its surrounding ions depends on the chemical

nature of the ions and on the concentration. Hydrogen electrodes are widely used for determining the effective concentration of hydrogen ions. See ELECTROCHEMISTRY.

**Colloids.** Colloids are very small particles which possess very large surface areas for a given weight. They adsorb ions, take on electrical charges, and acquire special properties which differ from those of true solutions and from systems which contain large particles. The study of colloids is a branch of physical chemistry which finds many applications in biology, and in industry. See COLLOID.

**Molecular structure.** Much valuable information concerning the chemical and physical properties of chemical compounds can be obtained from a knowledge of the molecular structure. The arrangement of atoms and electrons within the molecule and the attractive forces which hold them together can be determined from physical measurements of refractive index, rotation of polarized light, absorption of light of different wavelengths, diffraction of electron beams, and dielectric constant. Predictions of chemical behavior are made on the basis of the type of bonding. The electron pair, which holds atoms together, is particularly important in organic chemistry, and the electrostatic binding is particularly important in inorganic chemistry. Quantitative calculations are made on the basis of quantum mechanics.

Ordinary thermal chemical reactions are brought about by activating collisions between rapidly moving molecules. They can be brought about also by the absorption of light. The branch of physical chemistry which studies the relation between activation by light and chemical reaction is known as photochemistry. Chemical activation, produced by very high energy radiation such as x-rays or by radioactivity is studied in a branch of physical chemistry known as radiation chemistry. See CHEMISTRY MOLECULAR STRUCTURE AND SPECTRA; PHOTOCHEMISTRY. [F.D.]

**Bibliography:** F. Daniels and R. A. Alberty *Physical Chemistry*, 1955; S. Glasstone, *Textbook of Physical Chemistry*, 2d ed., 1946; C. F. Prutton and S. Maron, *Fundamental Principles of Physical Chemistry*, 1951.

## Physical law

A physical phenomenon is said to be controlled or governed by a physical law when the phenomenon is one of a broad class of phenomena such that it is possible to formulate some regularity which applies to all members of the class. If the phenomenon is one which can be described in terms of numerical measurement, the law is often formulated in terms of mathematical relations between the numbers obtained by measurement, as in the inverse square law of universal gravitation. However, a mathematical formulation of a natural law is by no means necessary.

It is implicit in the notion of physical law that there are no exceptions, and, unlike man-made law, it may not be "violated." [P.W.B.]

## Physical measurement

In science and engineering, and in much of everyday life, quantitative information is essential for coordination of activities and efficiency of communication. Time, distance, temperature, color, weight, volume, and hundreds of other physical quantities must be described in terms which have the same meaning to everyone. This is made possible by comparing the physical quantity, or another quantity functionally related to it, with a constant or reproducible quantity of the same kind and of a fixed size or magnitude defined by a standard. The magnitude embodied by this standard, or some division or function of it, is taken as a unit. The comparison is the process of measurement. The general area of scientific activity relating to standards for physical measurement is called metrology. The measured quantity may then be expressed by a number (the magnitude ratio) and the name of the unit, for example a length of 1.54 meters. See UNITS, SYSTEMS OF.

**Physical standards.** The U.S. national standards of weights and measures are based on, or related to, the internationally accepted standards for mass, length, time, and temperature. The standard for length is the wavelength of the orange light emitted when a gas consisting of the pure krypton isotope of mass number 86 is excited in an electrical discharge; the unit of length, the meter, is defined as 1,650,763.73 times this wavelength. The standard for mass is a platinum cylinder; its mass is taken as the unit, the kilogram. The standard for time is the cyclic motion of the Earth-Sun system; the unit of time, the second, is defined as  $1/31,556,925.9747$  of the "tropical year" 1900.

The standard of temperature is the temperature of water at its triple point (where ice, water, and water vapor are in equilibrium). The unit of temperature is the degree on the Kelvin (or thermodynamic) scale defined by taking the temperature of the triple point of water as  $273.16^\circ\text{K}$  with respect to the absolute zero of temperature (thermodynamically defined as the condition in which all kinetic energy has been abstracted from the molecules). This condition cannot be exactly attained, although it can be approached very closely. See ABSOLUTE ZERO; TRIPLE POINT.

The units of other physical quantities are defined in terms of these four units by appropriate defining equations. For example, force is defined as the product of mass and acceleration. Units of force, the dyne in the centimeter-gram-second (cgs) system and the newton in the meter-kilogram-second (mks) system, are defined by the following equations

$$1 \text{ dyne} = 1 \text{ gram} \times \frac{1 \text{ cm/1 sec}}{1 \text{ sec}}$$

$$1 \text{ newton} = 1 \text{ kg} \times \frac{1 \text{ m/1 sec}}{1 \text{ sec}}$$

As another example, viscosity is the force per unit area per unit velocity gradient. The unit of vis-

cosity, called the poise, is defined by the equation

$$1 \text{ poise} = \frac{1 \text{ dyne-sec}}{\text{cm}^2}$$

See DIMENSIONAL ANALYSIS.

To facilitate measurements of many of the other quantities, derived or calibration standards have been established which embody or reproduce accurately known or definite values of the quantities, such as standard cells for voltage, quartz oscillators for frequency, and pure liquids for viscosity.

The defining equations for the units of some other quantities have numerical coefficients, such as  $\pi/4$ ,  $1/2$ ,  $c$  (= speed of light), which reflect geometrical factors or physical principles.

By arbitrarily choosing different physical quantities on which to base the units in terms of which other quantities are to be defined, various other systems of units would be possible. However, the many difficulties in making experimental tests to establish useful standards would be little reduced, so there is no immediate prospect of international changing of these four quantities for which arbitrary, or "absolute," standards and units have been established.

It is possible, however, to change the standards for these quantities. For example, since the frequency of atomic vibrations is experimentally found to be much more nearly constant than is the frequency of the earth's rotation, the standard for time may someday be taken as a transition frequency of atoms of a particular isotope; this frequency can be determined precisely and conveniently. The unit of time, the second, would probably be unchanged in magnitude, but would be redefined in terms of the atomic frequency. See ATOMIC CLOCK.

The standard of length was redefined by international agreement in September, 1960, in terms of the wavelength of pure krypton, as specified above. Up to this time, the standard was a bar of platinum-iridium kept at the International Bureau of Weights and Measures in France. There are two advantages in redefining the standard of length in terms of wavelength: (1) the wavelength can be generated wherever needed, so that a primary (absolute) standard of length can be available in laboratories, or even in shops, to permit more accurate measurements, and (2) the chosen wavelength is presumably constant and unchanging, whereas the meter bar may be subject to slow changes in dimension with time and is, of course, subject to destruction, damage, or loss. See WAVELENGTH STANDARDS.

There appears to be no immediate practical way of realizing standards for mass from atomic properties that would permit precision of comparison as good as can be attained with present methods.

On July 1, 1959, by agreement among the national standards laboratories of the countries using English units in trade and industry, the slight differences between different yards and pounds were eliminated by defining the yard as 0.9144 meter, and the pound as 0.45359237 kilogram.



The corresponding value of the inch is now 2.54 centimeters, exactly. The old United States yard and inch had been about two parts per million larger than the new yard and inch.

**Accuracy of measurements.** In general, any measurement has less than perfect accuracy; there is some error or uncertainty as to the true, or exact, numerical ratio between the measured quantity and the unit. These errors are either observation or calibration errors. Observation errors are errors or uncertainties in reading scales and interpolating between scale divisions, or in synchronizing the readings with the events. Calibration errors are errors or uncertainties in the conversion factors relating the primary measured quantity and its representative quantity (such as when acceleration is measured in terms of the voltage generated in a piezoelectric material) or in the relation between the indicated readings and the units of the final measured quantity.

There are practical limits to the statistical exactness of comparison, even with repeated measurement. Furthermore, there are other sources of error due to inherent imperfections of materials and structures, which cause the phenomena of drift, lag, hysteresis, damping, and resonance.

Drift is the gradual change of instrument reading after a change to a different but constant value of the measured quantity.

Lag is failure of the instrument reading to follow changes in the measured quantity instantly or exactly. The time constant of an instrument is the time required for the indication to change by  $1/e$  of a sudden change in the measured quantity in response to this change. See **TIME CONSTANT**.

Hysteresis, which results from lag or drift, is the difference between readings of the measured quantity, for corresponding actual magnitudes of that quantity, when the quantity is increasing and when it is decreasing.

Damping is the result of frictional forces (either viscous or coulomb) which cause greater lag than that due to inertial effects alone. See **DAMPING**.

Resonance is the condition of oscillation which results when the rapidity of change of measured quantity is close to the natural rapidity of response of the instrument. See **RESONANCE (ACOUSTICS AND MECHANICS)**.

The change of properties of materials or structures with temperature, pressure, humidity, radiation, vibration, or other environmental conditions may also cause errors. Thus, the analysis of experiments to detect all possible sources of error, the design of experimental procedures to minimize them, and the development of mathematical techniques for estimating their probable magnitudes constitute an important part of any measurement in which precision is important.

**Measurement techniques.** The comparison of quantities as to equality, the counting of units, and the determination of the coincidence of events—all involved in physical measurement—may be done by an observer (usually using his visual, aural, or tac-

tile faculties) or by instruments which display or record the results, or apply them to automatic computation or control. See **INSTRUMENTATION**.

Direct measurement involves comparison with a standard (such as a meter bar) or measurement by a calibrated instrument (such as a voltmeter). Indirect measurements are those derived from measurements of related quantities, for example, the mass of the electron can be derived from measurements on the bending of the electron path in a magnetic field and separate measurements of its charge. Absolute measurements are those derived from measurements of the primary quantities involved; as in determination of acceleration from length and time intervals, rather than from reaction force.

Once a standard of any physical quantity is established (by definition or absolute measurement), any instrument of known stability, calibrated by measuring the standard, or multiples or divisions thereof, may be used as a standard instrument to calibrate other instruments by comparison of readings when measuring constant or repeatable values of the quantity.

Space permits mention of only a few special techniques devised to reduce uncertainties and errors in measurement:

1. In weighing on a balance, known weights are substituted for unknown weights previously balanced by counterweights, so that such uncertainties as knife-edge placement and beam lengths do not affect the comparison.

2. The effect of a quantity to be measured may be nearly offset by a similar known and fixed quantity, and the difference measured by an instrument of lesser range and of correspondingly higher precision.

3. If the standard offsetting quantity can be divided into sufficiently small or continuously variable fractions, it can be adjusted for equality to within the smallest limits detectable. This is the so-called null, or balancing, method of measurement.

4. A number of small quantities of the same kind may be combined so that the cumulative magnitude is appropriate for measurement with available methods.

5. The measurand may be time-modulated to permit better discrimination between the measurand and extraneous effects on the measuring instrument.

6. The undesired effects of environmental factors, such as temperature and external magnetic fields, may be compensated for in the instrument system by elements which are responsive to the disturbing factor and interact with the measuring or indicating means to offset such effects.

7. A series of observations of a particular measurand permits statistical determination of an average and leads to increased precision; also an estimate of the probable error of observation (not including systematic errors) can be obtained.

8. Based on careful design of experiments, observations may be made under a wide variety of



conditions, with controlled variation in the factors considered as possible sources of error, permitting statistical estimation of the magnitude of the various errors and appropriate correction for them.

[W.A.W.]

*Bibliography:* See INSTRUMENTATION.

## Physical science

The fields of inquiry to which the general designation science may be appropriately applied are broadly divided into social science and natural science. The latter is further subdivided into biology and physical science. Physical science is generally considered to include astronomy, chemistry, geology, mineralogy, meteorology, and physics. These overlap more or less, as illustrated by astrophysics, chemical physics, physical chemistry, and geophysics. There is overlap, likewise, between the physical and biological sciences, as seen in biochemistry, biophysics, virology, and the close relation between geology and paleontology. The boundaries implied in all such classifications are artificial and consist of regions where one field shades into another.

Chemistry and physics differ from astronomy, meteorology, and geology in that they are concerned with the properties of matter and energy encountered alike upon and within the earth, the planets, and stars. For this reason, chemistry and physics are not set apart from, but rather pervade the other sciences.

To regard the several areas of scientific inquiry as separated by sharp definable boundaries is unrealistic. Cross-fertilization has produced some of the most notable advances in science, and an artificial barrier can advantageously be accepted as a challenge by a scientist with an adventurous mind. See SCIENCE.

[J.H.H.]

## Physics

The original objective of physics, formerly called natural philosophy, was to understand the structure of the natural world and explain natural phenomena. In time, various specialized sciences broke away from physics to form autonomous fields of investigation. In this process physics retained its original aim of concerning itself with those aspects of nature which could be understood in a fundamental way in terms of elementary principles and laws.

**Basic parts.** The most basic parts of physics are mechanics and field theory. Mechanics is concerned with the motion of particles or bodies under the action of given forces. The physics of fields is concerned with the origin, nature, and properties of gravitational, electromagnetic, nuclear, and other force fields. Taken together, mechanics and field theory constitute the most fundamental approach to an understanding of natural phenomena which science offers. The ultimate aim is to understand all natural phenomena in these terms. See FIELD THEORY, CLASSICAL; MECHANICS; QUANTUM FIELD THEORY.

The older, or classical, divisions of physics were based on certain general classes of natural phenomena to which the methods of physics had been found particularly applicable. These consisted of classical mechanics with branches in celestial mechanics, hydrodynamics, and ballistics; heat and thermodynamics; kinetic theory of gases and statistical mechanics; optics; acoustics; and electricity and electromagnetism. These divisions are all still current, but in present usage many of them tend more and more to designate branches of applied physics or technology, and less and less inherent divisions in physics itself.

**Branches of modern physics.** The divisions or branches of modern physics are made in accordance with particular types of structures in nature with which each branch is concerned. Thus elementary-particle or ultra-high-energy physics is the most recent branch and is concerned with understanding the properties and behavior of elementary particles as such, and more particularly of the heavy unstable particles—mesons, hyperons, and antiparticles—which are produced in collisions involving energies above about 150,000,000 electron volts. The next branch in this classification is nuclear physics, which is concerned with associations of neutrons and protons forming the nuclei of atoms; their structure, properties, and energy states; reactions between nuclei, including scattering processes and radioactivity; and related phenomena such as the interaction of high-speed nuclear particles with matter. Atomic physics is concerned with the structure and properties of atoms as determined by the electrons outside the nucleus; the states of motion of these electrons including such topics as energy levels, angular momentum properties, and magnetic moments; and the absorption and emission of radiation by atoms.

Continuing with this classification in ascending complexity there is molecular physics, which is concerned with systems of atoms formed into molecules, the nature of intermolecular forces, chemical binding, vibration and rotation spectra of molecules, and the like. Next in order are solid-state physics, physics of liquids, physics of gases, and more recently plasma physics, which deals with properties of highly ionized atoms forming a mixture of bare nuclei and electrons called an ion plasma.

In this same classification could also be included biophysics, which deals with the application of physical methods and types of explanation to biological systems and structures.

Other more specialized classifications may be made in accordance with particular instruments or techniques such as x-ray diffraction, neutron diffraction, mass spectrometry, infrared spectroscopy, and seismology. The special field of low-temperature physics is characterized not only by special instruments involved in the production and measurement of low temperatures in the range of liquid helium but also by the phenomena of superconductivity and superfluidity which occur only in this temperature

range. Other fields, such as astrophysics and geophysics, are concerned with aspects of other sciences to which physics is applicable.

**Mathematical physics** is the study of physical phenomena by means of mathematics, and includes the more mathematical parts of all branches of physics, as well as most of the content of statistical mechanics, quantum mechanics, relativity, and field theory. A distinction is often made between mathematical physics and theoretical physics, in which the latter, although still entirely mathematical in form, is thought of as being more closely related to experimental physics. Neither mathematical nor theoretical physics can really be separated from experimental physics, since a complete understanding of nature can only be obtained by the application of both theory and experiment.

**Aim of physics.** In every area physics is characterized not so much by its subject matter content as by the precision and depth of understanding which it seeks. The aim of physics is the construction of a unified theoretical scheme in mathematical terms whose structure and behavior duplicates that of the whole natural world in the most comprehensive manner possible. Where other sciences are content to describe and relate phenomena in terms of restricted concepts peculiar to their own disciplines, physics always seeks to understand the same phenomena as a special manifestation of the underlying uniform structure of nature as a whole. In line with this objective, physics is characterized by accurate instrumentation, precision of measurement, and the expression of its results in mathematical terms.

For the major areas of physics and for additional listings of articles in physics, see ACOUSTICS; ASTROPHYSICS; ATOMIC PHYSICS; BIOPHYSICS; ELECTRICITY; ELECTROMAGNETISM; HEAT; LOW-TEMPERATURE PHYSICS; MECHANICS, CLASSICAL; MOLECULAR PHYSICS; NUCLEAR PHYSICS; OPTICS; SOLID-STATE PHYSICS; THEORETICAL PHYSICS.

[W.G.P.]

## Physiological acoustics

A term used to refer to the physiology involved in the process of speaking and hearing. Physiological acoustics includes the action of the larynx, glottis, throat, mouth, tongue, and teeth in the process of speaking. It includes the action of the eardrum, the small bones of the middle ear, the inner ear consisting of the oval window, cochlea, basilar membrane, and round window, and also the action of the nerves carrying the acoustic stimulation to the brain in the process of hearing. A study of the effect of various kinds of sounds upon all the physiological processes in the body is also included in this field. The terms physiological acoustics and psychoacoustics are sometimes used interchangeably. Technically, however, only the topics mentioned here fall under the domain of physiological acoustics. See HEARING; SPEECH; *see also* PSYCHOACOUSTICS.

[H.F.L.]

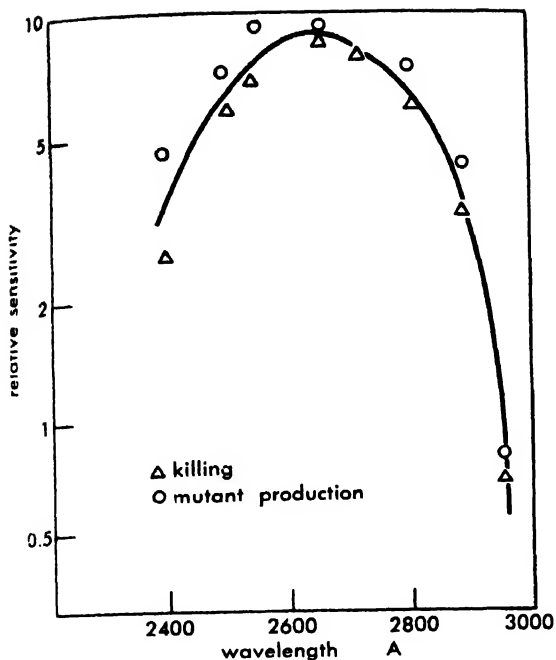
## Physiological action spectra

An action spectrum is a representation of the comparative effects of different wavelengths of light on living systems or on the components of living systems. A knowledge of the effects of different wavelengths on living systems helps lead to an understanding of the detailed mechanisms of energy transfer and utilization and to a determination of the essential compounds involved in light action on living systems. The work of action spectroscopy is founded on the firm bases that energy must be absorbed before it is utilized and that each chemical compound has a characteristic absorption spectrum. Therefore, the shape of the action spectrum may lead to the identification of the absorbing molecules. Action spectroscopy has been used extensively to study three classes of compounds: porphyrin-containing proteins, nucleic acid polymers, and plant pigments. See ABSORPTION (ELECTROMAGNETIC RADIATION).

**Beneficial effects of light.** Proteins which contain porphyrin prosthetic groups, such as hemoglobin and cytochrome, play an important part in the transport of oxygen and in the oxidation-reduction systems of cells. Such prosthetic groups have a high affinity for carbon monoxide (CO) and therefore the respiration of living systems is inhibited by carbon monoxide. Light can remove this carbon monoxide inhibition by causing the dissociation of the porphyrin-CO complex. The extensive studies of O. H. Warburg and his collaborators on the action spectra for the removal of the carbon monoxide inhibition respiration led to the identification of the respiratory enzyme as a porphyrin-containing protein. The fact that light absorbed not only in the porphyrin ring but also in the protein can cause the dissociation of CO indicates that energy may be transferred over distances of about 30 angstrom units (Å). See CYTOCHROME; HEMOGLOBIN; PORPHYRIN.

**Destructive effects of light.** The beneficial effect of light mentioned above is to be contrasted with the destructive effect of light in the ultraviolet spectral region. This light can destroy the enzymatic activity of proteins and can kill bacteria and viruses. Action spectra for the killing and the production of mutants in yeast, shown in the accompanying figure, indicate that the most sensitive wavelengths are in the neighborhood of 2650 Å. The shape of the action spectrum is very similar to the absorption spectrum of nucleic acids and indicates that nucleic acid polymers are essential components in the duplication of living systems. Some of the effects of ultraviolet light on deoxyribonucleic acid (DNA) are reversed when the system is illuminated by intense visible light. Such reversibility is known as photoreversal or photoreactivation. See BACTERIA; NUCLEIC ACID; PROTEIN; ULTRAVIOLET RADIATION (BIOLOGY).

**Photosynthesis.** Because the sun is the most important source of energy, it is not surprising that



The action spectra for the killing of yeast and the production of mutants in yeast by ultraviolet light. (After data of C. Raut and W. L. Simpson, *Arch. Biochem. Biophys.*, 57(1):218-228, 1955)

Photosynthesis is one of the most thoroughly studied of the fields which concern the energy exchanges between light and living systems. The identification of chlorophyll as an essential component of photosynthesis is made by comparing the action spectrum for photosynthesis with the absorption spectrum of chlorophyll. The similarity between the two indicates the role that chlorophyll plays in photosynthesis. On the other hand, the action spectra for photosynthesis in some algae do not resemble the absorption spectrum of chlorophyll. In these cases, light is absorbed in other pigment molecules and the energy is transferred to chlorophyll to be used for photosynthesis. Action spectra have shown the existence of wavelength-dependent effects on such diverse photoperiodic properties of living things as the germination of seeds, the flowering of plants, and the change in coloration of animals with the seasons. All these effects are governed by an elaborate interplay between red and far red light.

**Mechanism.** The primary process in light action is the absorption of a light quantum by a molecule. The molecule is thus raised to an excited state. The excitation energy may be passed on to other molecules and utilized for chemical reaction as in photosynthesis, or the extra energy may be used to break bonds and alter chemical structure as in the inactivation of bacteria by ultraviolet light or the excitation energy may be reemitted as fluorescent light or ultimately degraded as heat. In the latter case there may be no effect of the absorbed radiation at all. A particularly interesting feature

of excited states is the ability of energy to be transferred over appreciable distances from one molecule to another by a process known as energy transfer. Such processes permit energy absorbed in one molecule to appear in a different type of molecule. The existence of resonance-energy transfers may be inferred from a comparison of the action absorption and the fluorescent spectra of the irradiated material. See CHLOROPHYLL; PHOTOPERIODISM IN PLANTS; PHOTOSYNTHESIS; VISION. [R.B.S.]

**Bibliography:** A. Hollaender (ed.), *Radiation Biology*, vol. 2, 1955, vol. 3, 1956; C. Reid, *Excited States in Chemistry and Biology*, 1957.

## Physiology, general

A division of physiological science in which the basic activities of living organisms that occur in most cells and tissues are studied by physical and chemical methods. Its living materials range from the simplest unicellular forms to man himself. Thus, oxidations and reductions, which are of universal occurrence in both plants and animals, are examples of such basic activities. Many cells accumulate certain ions within their cytoplasm while tending to exclude others. Ion gradients are therefore typically present across living membranes, associated with the development of electromotive forces (BIOELECTRIC MODEL). Numerous chemical reactions occurring within cells are speeded by organic catalysts, known as enzymes, whose mode of action is similar throughout the living world. The experimental study of such basic phenomena constitutes the field of general physiology.

In method and subject matter, general physiology differs considerably from mammalian physiology. The latter division of the science deals particularly with the functional activities of the most highly developed animals. It often becomes an adjunct of medical practice, with little consideration of basic principles. It tends to be "organ" physiology, which, as Ivan Pavlov once remarked, "has begun its study from the midst of life; the beginning, the basis of life, is in the cell." General physiology must also be distinguished from comparative physiology, in which emphasis is placed as much upon the differences between organisms as upon their similarities. However, no sharp distinction can be made between the three subdivisions of the subject. The well-trained physiologist draws his materials from all.

**Development.** Long before general physiology emerged as a separate discipline, observations were accumulating which properly belong to it. Thus, the generation of electricity by living organisms constitutes an important chapter in general physiological thought. Two species of electrical fish were known to the ancient Greeks and Egyptians. The Roman physicians used the shocks generated by *Torpedo* to treat various human diseases, thus inaugurating electroshock therapy. Interest in these animals was a potent factor in the develop-

ment of electrical science in the eighteenth century. Alessandro Volta called his voltaic pile an artificial *Torpedo*. See ELECTRIC ORGAN (BIOLOGY).

Physiological observations on living cells have led to other important advances in physical science. In 1827, Robert Brown reported on "the general existence of active molecules in organic and inorganic bodies." In examining the ovules and pollen grains of various plants under his microscope he observed small particles within them to be "in rapid oscillatory motion." A systematic search revealed the same phenomenon in the most diverse materials. In his honor, it came to be known as "Brownian movement." It was later recognized as due to the thermal agitation of molecules and ions in the cytoplasm and in other fluids. Their own movement is invisible, but their continual bombardment of particles suspended in them causes the visible oscillations. This thermal activity is basically responsible for the diffusion of materials along their concentration gradients, in solids, liquids, and gases.

There are several further examples of physiological experiments leading to physical discoveries. In 1748, J. Nollet discovered osmosis by observing that a pig's bladder, filled with alcohol and dipped into water, increases in volume as water moves into it. The experiments of H. De Vries, beginning in 1884, showed that, if a sugar solution of a certain molar concentration just produces plasmolysis of plant cells, the same effect can be produced by a lower concentration of sodium chloride or of other salts. J. Van't Hoff and S. Arrhenius shortly used this data to support their theory of electrolytic dissociation, arguing that each salt molecule, upon solution, splits into ions, both of which contribute to the osmotic pressure, whereas the sugar molecules remain undissociated.

Thus, general physiology is more than a study of the functional activities of plants and animals. Broadly conceived, it is a search for phenomena which are universal, occurring in both the living and the nonliving worlds, to the end that a more fundamental understanding of vital activities can be attained. See ELECTROPHYSIOLOGY (HEART).

**Development of general physiology.** General physiology was first established as a distinct discipline by the devoted labors of the great French physiologist, Claude Bernard (1813–1878), who introduced the term. In a series of brilliant investigations he added greatly to the subject matter of the science. His experimental studies included the mechanism of glandular secretion (particularly in the pancreas), animal glycogenesis, vasodilator and vasoconstrictor nerves, animal heat, anesthesia, poisons (especially curare and carbon monoxide), and a wide array of other subjects.

In 1865, Bernard wrote that "general physiology is the basic biological science toward which all others converge. Its problem is to determine the elementary condition of vital phenomena." Thinking more specifically about methodology, he argued that "the properties of living matter can be learned

only through their relation to the properties of inorganic matter; it follows that the biological sciences must have, as their necessary foundation, the physico-chemical sciences from which they borrow their means of analysis, and their methods of investigation."

Bernard's final review and summary was published in 1878. In it he developed his conception of an internal environment in the higher organism whose composition and physical state are so regulated, particularly by central nervous reflex action, that their cells live in a nearly constant fluid medium, distributed through their blood, lymph and intercellular fluids. His famous dictum "*la fixité du milieu intérieur est la condition de la vie libre, indépendante*" is usually translated "the constancy of the internal environment is the condition of the free life," but this rendering loses certain of the original force by omitting "independent."

At a later time, Walter Cannon introduced the term homeostasis to describe the regulation and adjustment of vital functions so that a steady state exists, not only in the blood and tissue fluids, but in other bodily mechanisms (see HOMEOSTASIS). J. Barcroft has more recently argued that the free life is to be thought of as a free mental life, since it is the cells of the nervous system, which, above those of all other tissues, require constancy in their milieu.

Such conceptions must properly be considered at the highest level in general physiological thinking. They are broad generalizations which permit better interpretation of whole series of vital phenomena in many organs. It is unwise to accept any definition of general physiology which make synonymous with cell physiology, as M. Verwoerd has done. The study of individual cells is an important part of its field, but integrative thinking about the whole organism is its even higher responsibility. Research and theory must proceed at all levels.

One further example of high-level thinking in general physiology may be considered. Lawrence J. Henderson in 1913 wrote persuasively concerning the "fitness of the environment." He argued that the physical and chemical characteristics of the inorganic world are peculiarly adapted to sustain life, and possibly caused its appearance in the first place. His argument may be briefly summarized as follows:

Enormous quantities of carbon, hydrogen, oxygen, partly in the form of water and carbonic acid, are apparently inevitable constituents of the atmospheres and crusts of the larger planets. The known compounds of carbon and hydrogen, and of carbon, hydrogen, and oxygen, far surpass in number the compounds of any other elements. "These elements, therefore, are uniquely fitted to be the stuff of which life is formed and of the environment in which it exists." Water has a high specific heat. Therefore in the ocean depths there are only minor fluctuations of temperature. Ocean water carries with it everywhere a constant mixture

salts and gases. Its hydrogen ion concentration is practically invariable, being maintained largely by a buffer system of bicarbonates and carbonic acid. It has furnished, therefore, an optimal environment for the appearance of living organisms.

The contributions of Rudolf Höber, long the leading German scholar in the field, and William Maddock Bayliss, best known English writer may be mentioned. Following in the Bernard tradition, Höber, in 1945, declared that "it is not only possible, but of importance, to anchor physiology even deeper in physical chemistry than was done previously." A similar emphasis recurs in most of the other published works, in which may be found discussions of diffusion, osmosis and osmotic pressure, colloidal state, water and electrolytes, hydrogen ion concentration, oxidation and reduction, enzyme action, and the effects of temperature, in addition to treatment of the nature of protoplasm, the permeability of membranes, nutrition, secretion, digestion, respiration and metabolism, excitation and inhibition, muscular contraction, and nerve conduction. Indeed, there is no segment of physiological study which may not be considered in the broad spirit of general physiology. The science is a method of approach and an attitude of mind, rather than a treatment of any particular topic, or special field. See OSMOREGULATORY MECHANISMS.

**Areas of general physiology.** Although general physiology is far more than cell physiology, and reaches its highest levels only in integrative thinking concerning the whole organism, a large segment of its literature has been devoted to studies of cell permeability, particularly of the outer, or plasma membrane. Familiar materials for such studies are large plant cells, the egg cells of invertebrates, and vertebrate red blood corpuscles. By a variety of methods, the penetration of many substances has been studied. No one factor can account for the observed penetrations. Molecules of some substances, for instance, water, oxygen, carbon dioxide, and urea, can enter and leave the cell because they are small and can readily make their way through small pores in the membranes. In many cases the ability of nonelectrolytes to penetrate is at least partly related to the lipoid solubility of the material, as E. Overton stressed in 1899. In the case of electrolytes, the electrical charge on the walls of the membrane pores seems to determine the passage. When the membrane charge is electronegative, as is usually the case, positively charged cations penetrate much more readily than do negatively charged anions. When the membrane charge is electropositive, as in the vertebrate red blood corpuscle, anions penetrate most readily. Since ions of one sign are preferentially excluded, electrolyte passage is relatively difficult, and the electrical resistance of living membranes is relatively high. In the usual case the larger organic molecules, such as fats or proteins, cannot enter living cells as such, although their constituent building stones, the fatty acids and amino acids, are able to do so. However, in certain special situations,

as in the cells of the liver, there is evidence that molecules of the plasma proteins can enter and leave with relative ease. See CELL (BIOLOGICAL).

**Models.** At an earlier time when much less was known about cell structure and function, there was a strong tendency among general physiologists to seek aid in their thinking by the construction of models. Inorganic or nonliving organic systems may be built to imitate some vital activity, although they are never able to reproduce all of its aspects. Their construction emphasizes the desire of investigators in this field to deal with universal phenomena. A summary description of many of these models was given by T. C. Barnes in 1937. They include 1. Traube's model of the cell, conceived in 1867, which was formed by letting drops of gelatin fall into 5% tannic acid, which causes the formation of a precipitation membrane around the gelatin. These artificial cell membranes are sufficiently dense to prevent the permeation of glucose. If glucose is added to the gelatin before the cells are formed, it later causes them to swell because of osmotic movement of water into them. E. N. Harvey in 1912 formed artificial cells by adding egg white to water, then shaking with chloroform until protein-coated chloroform droplets appeared. If lecithin is added to the chloroform and the chloroform droplets are allowed to evaporate in water, protein-covered lecithin cells remain, which resemble sea urchin eggs in many properties.

Traube is also responsible for the introduction of a model of the cell membrane which has been widely employed in physicochemical studies. When a solution of potassium ferrocyanide comes into contact with one of copper sulfate, a membrane of copper ferrocyanide is formed, which may be supported in a porous porcelain cup. It permits the passage of water and a few other small molecules but excludes sugar and larger molecules. This membrane closely approximates the ideal semipermeable membrane, which is permeable to water but to no substance dissolved in the water. Living cell membranes are often called semipermeable, but it is more nearly correct to speak of them as selectively permeable.

Such model experiments have a certain pedagogical value, and may even aid in the planning of physiological experiments. But it is the vital activity which suggests the model, and not the reverse. The investigator should focus his attention upon the living cell or organism, to which his physical and chemical experimental methods should be directly applied.

**Use of the electron microscope.** In the last decade a great expansion of biological knowledge has occurred through the widespread use of the electron microscope which provides a wealth of new information concerning the ultrastructure of cells and tissues. The finer anatomy of cellular inclusions, such as the mitochondria and the Golgi apparatus, is being revealed. A system of fluid-filled tubes known as the endoplasmic reticulum has

been discovered. It is now known that many living membranes consist of two molecular sheets whose thickness can be measured. The pores imagined by general physiologists have now been shown to exist. The molecular architecture of nerve and muscle cells has been partially explored. The intricate ultrastructure of the neuromyal junction has been revealed. See NEUROPHYSIOLOGY.

At the cellular and subcellular levels, general physiology is thus presented with new opportunities and a great challenge. Older techniques are inadequate to study structures which can be seen only in ultrathin sections, and which in the course of their preparation for electromicroscopy become completely dry. New approaches to basic problems must be developed, combining physical, chemical, and histological techniques in a concerted effort to obtain still deeper insights into vital mechanisms. See ANIMAL GROWTH; CIRCULATION; EXCRETION; METABOLISM. [W.R.A.]

**Bibliography:** W. M. Bayliss, *Principles of General Physiology*, 4th ed., 1927; C. Bernard, *An Introduction to the Study of Experimental Medicine* (1865), H. C. Greene's translation, 1927; H. Davson, *A Textbook of General Physiology*, 1951; L. V. Heilbrunn, *Dynamics of Living Protoplasm*, 1956; L. V. Heilbrunn, *General Physiology*, 3d ed., 1952; R. Höber, et al., *Physical Chemistry of Cells and Tissues*, 1945; B. T. Scheer, *General Physiology*, 1953.

## Phytomastigophorea

A class of the subphylum Flagellata also known as the Phytomastigina. These are the plant flagellates which contain chlorophyll and other pigments, but also include colorless forms. Grass green is the usually observed color, primarily because the green flagellates are the largest. Those containing an excess of yellow pigments generally are smaller in size, and fewer species have unusual colors such as blue or red. Holophytic, saprophytic, and holozoic modes of nutrition occur, and specific chemical components may be demanded by individual species within the group.

Encystment is frequent among phytoflagellates, cyst composition being one method of determining relationships for some colorless species. Reproduction may occur within the cyst, or while the organism is active. Gamete formation is largely restricted to Phytomonadina, but life cycles may include an alternation of flagellate with palmella, or with ameboid generations. The Phytomastigophorea include six groups usually considered to be orders. These are the Chrysomonadida, Cryptomonadida, Phytomonadida, Euglenida, Chloromonadida, and Dinoflagellida. See articles on these groups. See also MASTIGOPHORA. [J.B.L.]

## Phytomonadida

An order of the class Phytomastigophorea. The protozoans, also known as the Phytomonadina, are grass-green, but a few are colorless (*Polytoma*). Individual cells may be as small as 8  $\mu$ . They closely

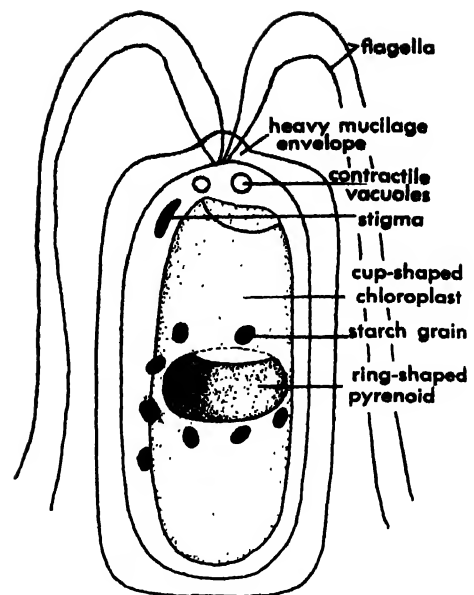


Fig. 1. *Carteria*, near *taticra*.

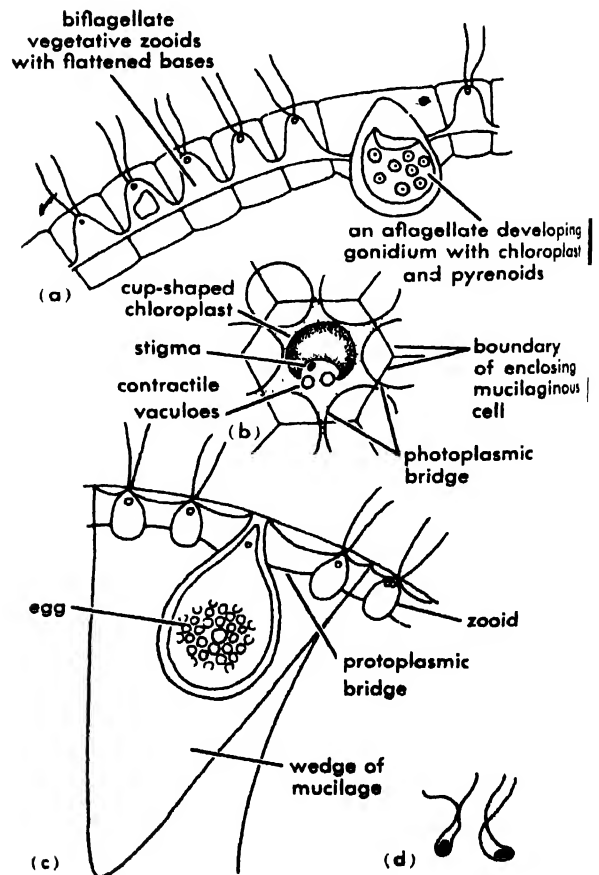


Fig. 2. (a) *Volvox globator*. Sectional view. (b) *V. globator*. Surface view of a single zooid. (c) *V. aureus*. Sectional view. The zooids are pear-shaped and are attached by threadlike protoplasmic strands to each other. The mucilage draws away from the thick outer boundary, but several cells are incorporated into deep wedge of mucilage. (d) *V. aureus*. Sperm, enlarged about 8 times by comparison with (c).



relate the flagellates to the algae. They have one flagellum (*Pedimonas*), two (*Chlamydomonas*), four (*Carteria*) (Fig. 1), or eight (*Polyblepharides*). Plural flagella are usually equal. The group is large and about one-fourth of the approximately 100 genera form palmelloid or dendroid colonies (*Tetraspora*, *Chlorangium*), with flagella only in reproductive cells. Cell walls are of cellulose and often they are thick. Chromatophores contain the same chlorophylls as higher plants. Pyrenoids have the usual form commonly found with starch as the reserve material.

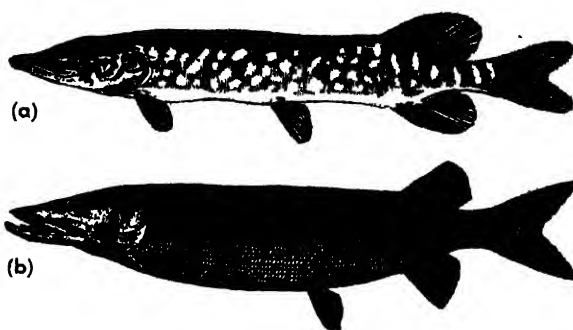
Unicellular species predominate, but at least 16 genera form colonies. These are irregular, linear, flat, or spherical. *Volvox globator* is a hollow sphere often 1 mm in diameter. Its mucilaginous periphery contains thousands of biflagellate zooids. Some zooids are gonidia that form new colonies asexually by a complicated multiple division. Isogamy, the production of like gametes which fuse, is common in Phytomonadida, but *Volvox globator* is monoecious, producing large eggs from certain zooids and packets of small sperm from others. Eggs, liberated into the hollow sphere, are fertilized by sperm. Thick rough walls envelop the fertilized eggs which eventually produce new colonies. Thus, *V. globator* illustrates sexual reproduction comparable to that of higher animals. Most phytomonads are fresh-water inhabitants; some are terrestrial, some marine. See PHYTOMASTIGOPHOREA; REPRODUCTION, ANIMAL. [J.B.L.]

### Piciformes

An order of birds containing six families, which have in common a peculiar arrangement of the tendons of the toes and other anatomical characters. All families nest in holes, and the young birds spend a relatively long period in the nest. The largest family is the Picidae, the woodpeckers, world-wide in distribution except for Madagascar, Australia, and most of Oceania. Woodpeckers show many adaptive modifications related to their climbing and feeding habits, notably the exceptionally long tongue and hyoid mechanism. Three families are confined to the Neotropical region: the jacanars (Galbulidae), puff birds (Bucconidae), and toucans (Ramphastidae). The barbets (Capitonidae) are pantropical, whereas the honey guides (Indicatoridae) are confined to the African and Indian regions. The honey guides are noted for their habit of leading man and other animals to bees' nests; a microorganism in the digestive tract permits the birds to digest beeswax. Honey guides are parasitic. They lay eggs in the hole-nests of such birds as barbets, woodpeckers, and starlings. See [K.C.P.]

### Pickrel

A name applied somewhat interchangeably to five of the six species of fishes in the genus *Esox*. Another commonly used name for these species is pike. The fourth species is the musky, or muskellunge, *Esox masquinongy*. All belong to the family Esocidae.



(a) The northern pike, *Esox lucius*; length to over 4 ft. (b) The muskellunge, *Esox masquinongy*; length to 8 ft. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

The pikes are Holarctic in their distribution. One species, *Esox reicheri*, occurs only in Siberia. The northern pike, *E. lucius*, found in the northern parts of North America and in Eurasia, is one of the finest sport fishes, striking with ferocity and fighting with determination when hooked. It frequently weighs 15 lb or more. Like all the *Esox* large enough to eat, it is an excellent food fish. There is a substantial commercial fishery for this species, primarily in the Canadian lakes.

The muskellunge is primarily a fish of the Great Lakes drainage system. There is a subspecies in the Ohio River system. Sixty-lb muskies are taken almost every year, and 40-lb fish are fairly common. This is the acknowledged king of American fresh-water sport fishes.

The chain pickerel, *Esox niger*, is an eastern species, fairly common in weedy water from Maine to Florida. It also occurs in the eastern part of the Great Lakes system, and northward up the Mississippi River into Missouri. This is a smaller fish than the northern pike, reaching a length of only 2 ft. The banded pickerel, *E. americanus*, occurs east of the Alleghenies. It is seldom over 1 ft in length. The smallest of all is the mud pickerel, *E. vermiculatus*, seldom reaching 1 ft in length and common in the Mississippi, Ohio, and Great Lakes systems.

All pike are voracious predators, recognized by the duck-bill head, the cylindrical body, and the small, posteriorly set dorsal and anal fins. See CLUPEIFORMES. [J.D.B.]

### Picrate

One of two types of compound which are derived from picric acid, 2,4,6-trinitrophenol. The hydrogen in the —OH group of the phenol is sufficiently labile to act as an acid and form salts with inorganic bases in the usual manner (for example, sodium picrate).

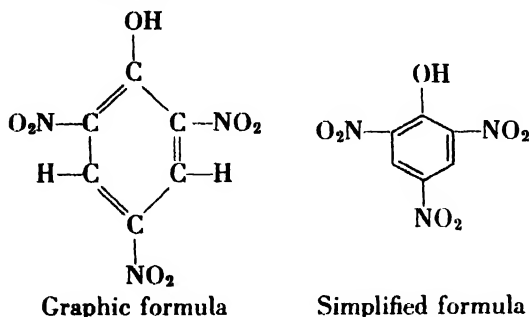
The second class of picrate is a group of molecular complexes formed when picric acid and aromatic compounds, such as naphthalene, are allowed to react. Although the nature of these complexes is not entirely understood, they are useful in the identification of aromatic compounds, because the



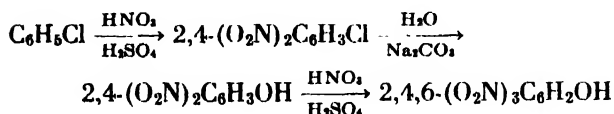
complexes are solids with characteristic melting points. These complexes are usually highly colored. See **PICRIC ACID**. [E.E.WR.]

### Picric acid

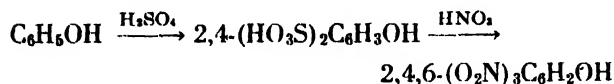
A phenol in which nitro ( $\text{NO}_2$ ) groups are present in the 2, 4, and 6 positions on the ring of carbon atoms. It approaches the mineral acids in acid



strength ( $\text{p}K_a = 0.80$ ). It is produced (1) by nitration of chlorobenzene to the dinitro derivative, hydrolysis of the latter, followed by another nitration;



or (2) by sulfonation of phenol, followed by nitration.



It has been used in treatment of burns, as an explosive, and as a yellow dye. See **NITROBENZENE**; **PHENOL**. [R.B.C.]

### Picrite

The term picrite has been used with several different meanings. It is generally considered to include certain medium- to fine-grained igneous rocks composed chiefly of olivine with smaller amounts of pyroxene, hornblende, and plagioclase feldspar (labradorite).

The feldspar content is slightly higher than that of peridotite and lower than that of gabbro. Certain analcite-bearing types, associated with teschenite, have also been included under the term picrite. A characteristic feature is poikilitic texture in which large pyroxene or hornblende crystals enclose numerous small grains of olivine.

Picrite is rare and is found in small intrusives (sills and dikes). It may also occur in the lower portions of basaltic lava flows where olivine and pyroxene crystals have accumulated under the influence of gravity. See **GABBRO**; **IGNEOUS ROCKS**; **PERIDOTITE**. [C.A.CA.]

### Pictorial drawing

A pictorial drawing shows a view of an object (actual or imagined) as it would be seen by an observer who looks at the object either in a chosen

direction or from a selected point of view. One such view usually suffices to give the reader a clear picture of the shape and details of the object. Pictorial sketches often are more readily made and more clearly understood than are front, top, and side views of an object (see **DESCRIPTIVE GEOMETRY**; **ENGINEERING DRAWING**). Pictorial drawings, either sketched freehand or made with drawing instruments, are frequently used by engineers and archi-

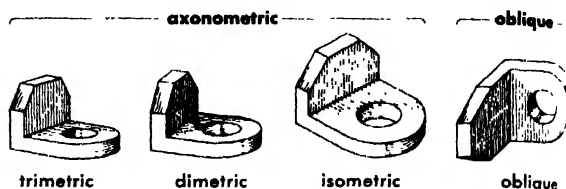


Fig. 1: Four different forms of pictorial drawing (From T. E. French and C. J. Vierck, *Graphic Science* McGraw-Hill, 1958)

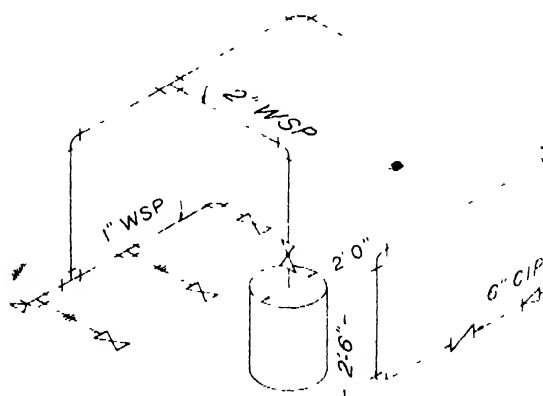


Fig. 2. Isometric piping drawing.

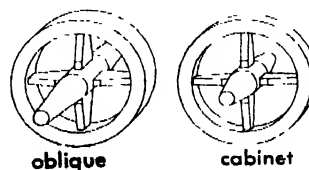


Fig. 3. Oblique and cabinet drawing. (From T. E. French and C. J. Vierck, *Graphic Science*, McGraw-Hill, 1958)

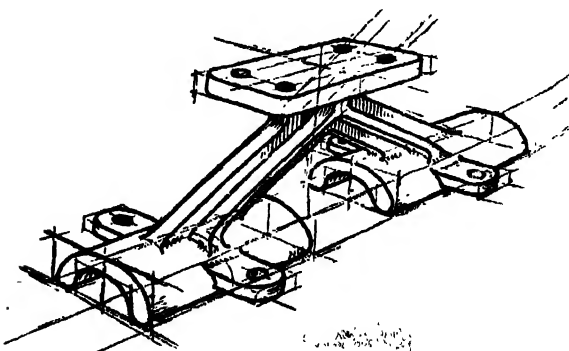


Fig. 4. A perspective sketch. (From T. E. French and C. J. Vierck, *A Manual of Engineering Drawing for Students and Draftsmen*, 8th ed., McGraw-Hill, 1958)

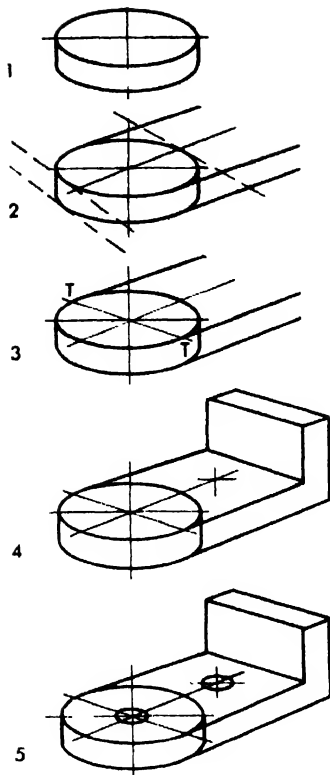


Fig. 5. Orientation of an ellipse to depict a circle.  
(From G. J. Hood and A. S. Palmerlee, *Geometry of Engineering Drawing*, 4th ed., McGraw-Hill, 1958)

tests to convey ideas to their assistants and clients. In making a pictorial drawing, it is important to select the viewing direction that shows the object and its details to the best advantage. The resultant drawing is orthographic if the viewing rays are considered as parallel, or perspective if the rays are considered as meeting at the eye of the observer. Making perspective drawings with instruments is time-consuming and requires considerable knowledge and skill.

Several types of nonperspective pictorial views can be sketched, or drawn with instruments (Fig. 1). The dimetric and trimetric drawings provide quite satisfactory views of the object, while the isometric view is not quite so satisfactory because of the distortion created by the high viewing angle. However, isometric drawings are relatively easy to make with drawing instruments. They are particularly useful for showing piping layouts (Fig. 2).

Cabinet and other oblique drawings, while not true orthographic views, offer a convenient method for drawing circles and other curves in their true shape (Fig. 3).

An effective pictorial sketch of an object can be made if proper attention is given to viewing direction, proportions, orientation of ellipses, and location of tangent points (Fig. 4). An essential feature in pictorial drawing is the proper orientation of the major (longest) axes of ellipses representing circles. The major axis of an ellipse should be

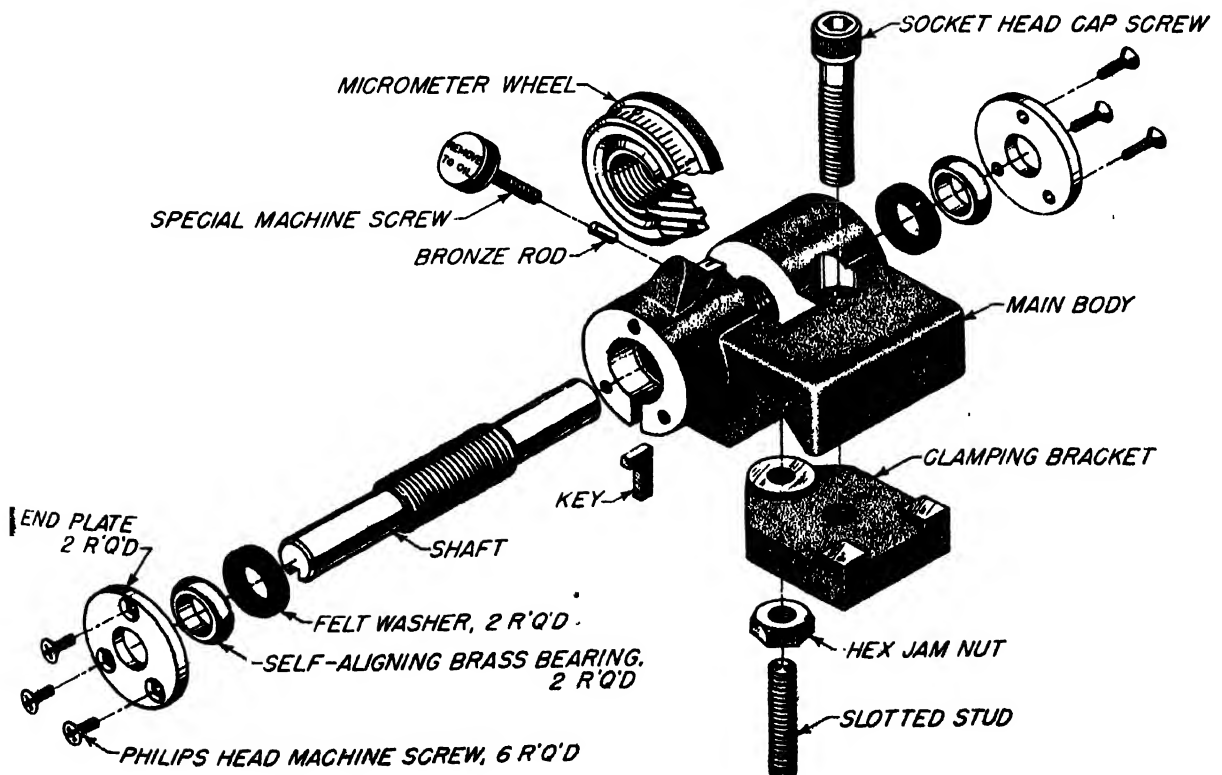


Fig. 6. Exploded-view production illustration. (From E. Zozzora, *Engineering Drawing*, 2d ed., McGraw-Hill, 1958)

drawn perpendicular to the axis of the circle (Fig. 5). The major axis of an ellipse representing a horizontal circle is horizontal, because it is perpendicular to the vertical axis of the circle.

Shaded exploded-view production illustrations greatly facilitate the learning process in assembly of machines and devices (Fig. 6). When this type of illustration is used, the initial assembly of parts into a machine has been found to be three or four times faster than if a conventional assembly drawing is used. [A.S.P.; C.J.B.]

## Pier

A wharf projecting perpendicularly or obliquely from shore, serving as berths for ships loading and discharging passengers and cargo. Construction usually takes the form of a pile-supported platform using steel, timber, and concrete materials. Sometimes quay wall or bulkhead construction is used around the periphery with earth fill inside. Piers may be open-deck or housed over with sheds. They may also have special cargo-handling equipment thereon, such as that used for loading iron ore. Uses of piers sometime extend to recreational purposes, such as fishing piers, or to community purposes, such as car parking. Spaces between adjacent piers are called slips. See COASTAL ENGINEERING; WHARF. [E.J.Q.]

## Piezoelectric crystal

A crystalline substance which exhibits the piezoelectric effect. This "pressure electricity" was first positively identified by the Curies in 1880, when they discovered that some crystals produced electric charges on parts of their surface when the crystals were compressed in particular directions, the charge disappearing when the pressure was removed. It was later discovered that these crystals become strained when subjected to electric fields; the piezoelectric deformation is directly proportional to the field and it reverses in sign as the sign of the field is reversed. These basic properties of piezoelectric crystals are used in electromechanical transducers, such as ultrasonic generators, microphones, phonograph pickups, and electromechanical resonators, such as the frequency-controlling quartz crystals. See PIEZOELECTRICITY.

**Piezoelectric materials.** The principal piezoelectric materials used commercially are crystalline quartz and Rochelle salt, although the latter is being superseded by other materials, such as barium titanate. Quartz has the important qualities of being a completely oxidized compound (silicon dioxide), and is almost insoluble in water. Therefore, it is chemically stable against changes occurring with time. It also has low internal losses when used as a vibrator. Rochelle salt has a large piezoelectric effect, and is thus useful in acoustical and vibrational devices where sensitivity is necessary, but it decomposes at high temperatures (55°C) and requires protection against moisture. Barium titanate provides lower sensitivity, but greater immunity to temperature and humidity effects. Other

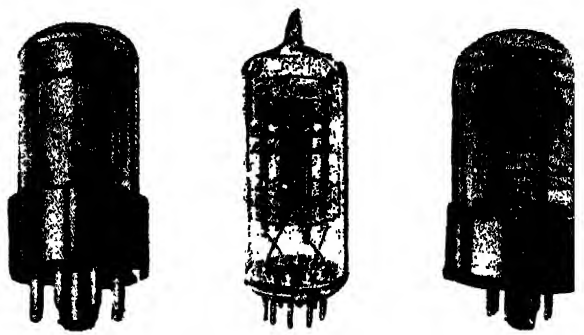


Fig. 1. Typical vacuum-mounted quartz resonators. (Northern Engineering Laboratories)

crystals that have been used for piezoelectric devices include tourmaline, ammonium dihydrogen phosphate (ADP), and ethylenediamine tartrate (EDT). See BARIUM TITANATE; QUARTZ; ROCHELLE SALT.

The quartz crystal resonator is the most important class of piezoelectric device. Its principal application is in the fields of frequency control and electric-wave filters. It is also used in transducers, especially where heat or moisture are factors.

**Characteristics and manufacture.** The electrical properties of quartz crystals as circuit elements, including their temperature coefficient of frequency, motional inductance and capacitance, series resistance, and electrode or shunt capacitance are largely determined by the dimensions and angles of rotation of the resonator surfaces with respect to the crystal axes (see CRYSTAL OPTICS). In commercial practice, raw quartz crystals are oriented by means of x-ray goniometers, the required angles of rotation are measured on the mounting jig, and the required blanks (unfinished slabs) cut from the mother crystal by diamond-faced saws. The blanks are then ground to the required frequency using lapping techniques with gradual sizes of abrasives to obtain a smooth finish. Some high-quality crystals are given optically polished surfaces. It is common practice to etch the surface of the crystal with fluorine compounds to remove microscopic surface irregularities.

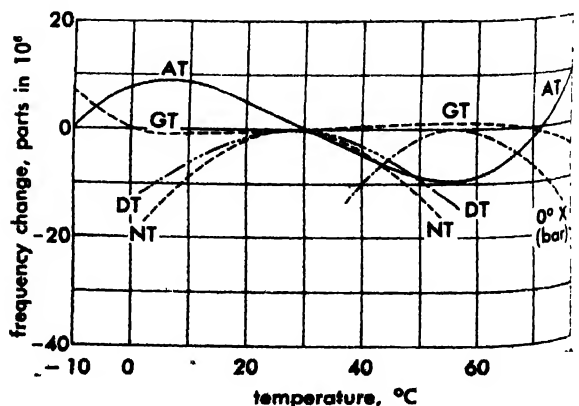


Fig. 2. Crystal frequency as a function of temperature for various crystal cuts.

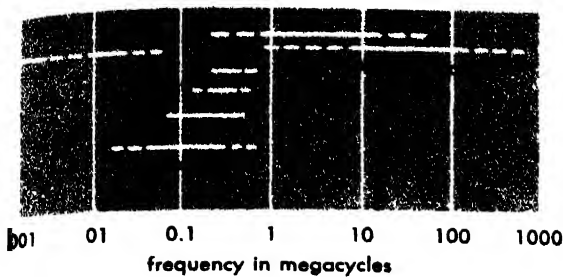


fig. 3. Frequency range of various crystal cuts.

Electrodes may be applied directly to the surface of the crystal, or they may be mounted externally in close proximity to the quartz element. Crystals with electrodes on their surfaces are frequently mounted by means of wires, which may also provide the connections to the electrodes. Containers for crystals may be hermetically sealed for protection from atmospheric effects, and some crystals are mounted in evacuated envelopes to improve  $Q$  and reduce aging drift. Examples of modern vacuum-mounted quartz crystals are shown in Fig. 1. Vacuum-mounted crystals are not capable of handling much power (usually less than 1 or 2 milliwatts maximum dissipation). Therefore, sealed crystals for power oscillators are usually mounted in an inert gas, such as nitrogen or helium.

The crystal-resonator cut chosen for a given application usually is dictated by the frequency at which the crystal must operate and the temperature range over which the crystal must work. A plot of typical frequency versus temperature curves for several different crystal cuts is shown in Fig. 2. A chart of frequency ranges covered by widely used crystal cuts is shown in Fig. 3.

**Applications.** Quartz crystal resonators are used for stabilizing the frequency of oscillators. The degree of stabilization depends on several factors, the principal ones being the  $Q$  of the resonator, the type of crystal cut used and temperature range of operation, the type of circuit used, and the amount of power dissipated in the resonator. Thermostatic elements are often used to enhance oscillator stability. The application of quartz crystals for oscillator stabilization has made possible our modern radio and television broadcasting industry and mobile radio communications with aircraft and ground vehicles. See **OSCILLATOR**.

Quartz crystal resonators are also used in electric-wave, or frequency-separation, filters. Many thousands of such crystals are used in telephone systems for carrier-frequency separation, and in radio communication equipment for selecting a desired signal frequency band while rejecting undesired frequencies. See **FILTER, ELECTRIC**.

Transducers using piezoelectric elements are used for converting vibrations into electrical signals, and are used in such applications as crystal microphones, phonograph pickups, vibration pickups, and dynamic pressure-sensing elements. The

inverse piezoelectric effect is used for converting electrical signals into mechanical vibrations. Thus, piezoelectric transducers are used in such applications as underwater sound ranging equipment (sonar, asdic), and in ultrasonic cleaning devices, which use a liquid medium for washing small to medium-sized objects. [F.D.L.]

**Bibliography:** W. G. Cady, *Piezoelectricity*, 1946; R. A. Heising, *Quartz Crystals for Electrical Circuits—Their Design and Manufacture*, 1946; W. P. Mason, *Piezo-electric Crystals and Their Application to Ultrasonics*, 1950; W. P. Mason, *Physical Acoustics and the Properties of Solids*, 1958; P. Vigoureux and C. F. Booth, *Quartz Vibrators and Their Applications*, 1950.

## Piezoelectricity

Electricity, or electric polarity, resulting from the application of mechanical pressure on a dielectric crystal. The application of a mechanical stress produces in certain dielectric (electrically nonconducting) crystals an electric polarization (electric dipole moment per cubic centimeter) which is proportional to this stress. See **POLARIZATION (DIELECTRICS)**. If the crystal is isolated, this polarization manifests itself as a voltage across the crystal, and if the crystal is shortcircuited, a flow of charge can be observed during loading. Conversely, application of a voltage between certain faces of the crystal produces a mechanical distortion of the material. This reciprocal relationship is referred to as the piezoelectric effect. The phenomenon of generation of a voltage under mechanical stress is referred to as the direct piezoelectric effect, and the mechanical strain produced in the crystal under electric stress is called the converse piezoelectric effect.

Piezoelectric materials are used extensively in transducers for converting a mechanical strain into an electrical signal. Such devices include microphones, phonograph pickups, vibration-sensing elements, and the like. The converse effect, in which a mechanical output is derived from an electrical signal input, is also widely used in such devices as sonic and ultrasonic transducers, headphones, loudspeakers, and cutting heads for disk recording. Both the direct and converse effects are employed in devices in which the mechanical resonance frequency of the crystal is of importance. Such devices include electric wave filters and frequency-control elements in electronic oscillator circuits. For additional information on applications, see **PIEZOELECTRIC CRYSTAL**; see also **DISK RECORDING**; **MICROPHONE**; **TRANSDUCER, UNDERWATER**; **ULTRASONICS**.

**Necessary condition.** The necessary condition for the piezoelectric effect is the absence of a center of symmetry in the crystal structure. Of the 32 crystal classes (see **CRYSTALLOGRAPHY**), 21 lack a center of symmetry, and with the exception of one class, all of these are piezoelectric. In the crystal class of lowest symmetry, any type of stress generates an electric polarization, whereas in crystals of

higher symmetry, only particular types of stress can produce a piezoelectric polarization. For a given crystal, the axis of polarization depends upon the type of the stress. There is no crystal class in which the piezoelectric polarization is confined to a single axis. In several crystal classes, however, it is confined to a plane. Hydrostatic pressure produces a piezoelectric polarization in the crystals of those 10 classes that show pyroelectricity in addition to piezoelectricity (see PYROELECTRICITY). The pyroelectric axis is then the axis of polarization.

The converse piezoelectric effect is a thermodynamic consequence of the direct piezoelectric effect. When a polarization  $P$  is induced in a piezoelectric crystal by an externally applied electric field  $E$ , the crystal suffers a small strain  $S$  which is proportional to the polarization  $P$ . In crystals with a normal dielectric behavior, the polarization  $P$  is proportional to the electric field  $E$ , and hence the strain is proportional to this field  $E$ . Superposed upon the piezoelectric strain  $S$  is a much smaller strain which is proportional to  $P^2$  (or  $E^2$ ). This strain is called the electrostrictive strain. It is present in any dielectric. See ELECTROSTRICTION.

**Matrix formulation.** The relation of the six components  $T_i$  of the stress tensor (three compressional components and three shear components) to the three components  $P_i$  of the polarization vector can be described by a scheme (matrix) of 18 piezo-

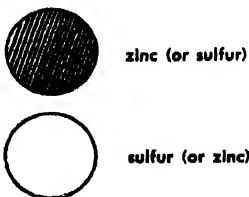
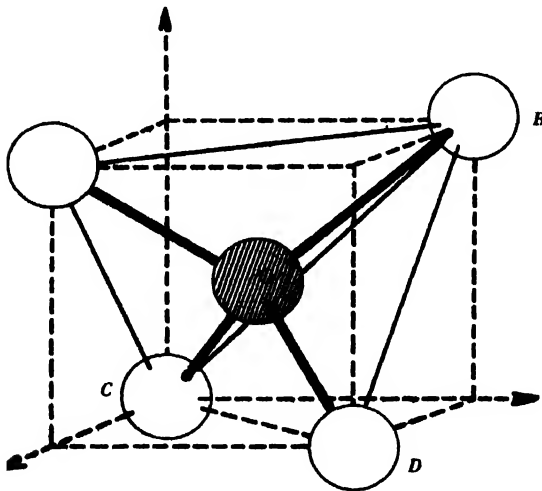


Fig. 1. Tetrahedral structure of zinc blende, ZnS. Only part of the unit cell is shown. Size of the circles has no relation to the size of the ions.

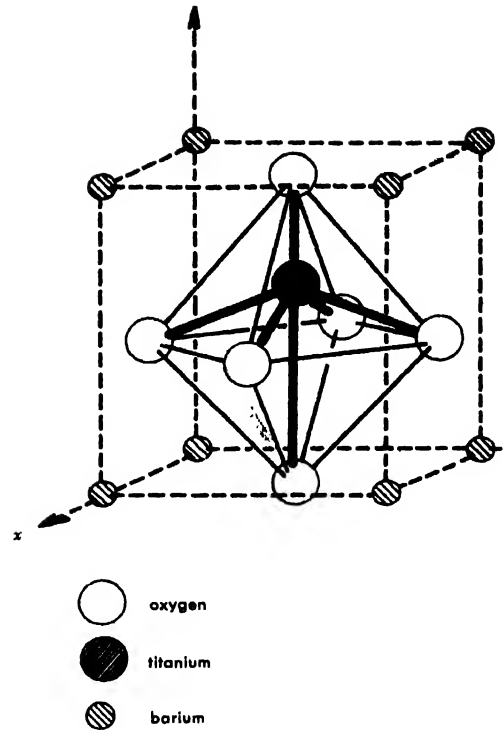


Fig. 2. Unit cell of tetragonal barium titanate, BaTiO<sub>3</sub>. Deviation from cubic symmetry is exaggerated. Size of the circles has no relation to the size of the ions.

electric moduli  $d_{ij}$ . The same scheme ( $d_{ij}$ ) also relates the three components  $E_i$  of the electric field to the six components  $S_j$  of the strain:

			Compression			Shear		
			$S_1$	$S_2$	$S_3$	$S_4$	$S_5$	$S_6$
			$T_1$	$T_2$	$T_3$	$T_4$	$T_5$	$T_6$
$E_1$	$P_1$		$d_{11}$	$d_{12}$	$d_{13}$	$d_{14}$	$d_{15}$	$d_{16}$
$E_2$	$P_2$		$d_{21}$	$d_{22}$	$d_{23}$	$d_{24}$	$d_{25}$	$d_{26}$
$E_3$	$P_3$		$d_{31}$	$d_{32}$	$d_{33}$	$d_{34}$	$d_{35}$	$d_{36}$

The direct effect is obtained by reading this scheme in rows:

$$P_i = - \sum_{j=1}^6 d_{ij} T_j \quad i = 1, 2, 3$$

The converse effect is obtained by reading it in columns:

$$S_j = \sum_{i=1}^3 d_{ij} E_i \quad j = 1, 2, \dots, 6$$

An analogous matrix ( $e_{ij}$ ) relates the strain to the polarization and the electric field to the stress:

$$P_i = \sum_{j=1}^6 e_{ij} S_j \quad i = 1, 2, 3$$

$$T_j = - \sum_{i=1}^3 e_{ij} E_i \quad j = 1, 2, \dots$$

The matrices ( $d_{ij}$ ) and ( $e_{ij}$ ) are not independent.

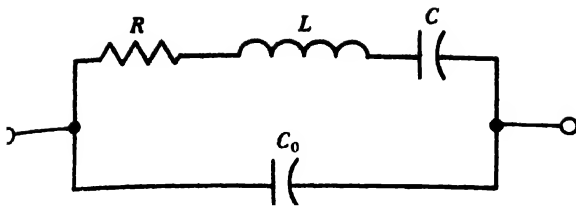


Fig. 3 Network equivalent to a piezoelectric resonator near and at a resonance frequency.

out are related by expressions involving the elasticity tensor  $c_{jkh}$  (for constant electric field  $E$ ):

$$e_{mh} = \sum_{j=1}^3 d_{mj} c_{jkh}$$

(see ELASTICITY). Alternative formulations are possible if one introduces the dielectric displacement  $D$  or visualizes the simultaneous action of electrical and mechanical stresses.

The number of independent matrix elements  $d_{ij}$  or  $e_{ij}$  depends upon the symmetry elements of the crystal. For the lowest symmetry, all 18 matrix elements are independent, whereas piezoelectric losses of higher symmetry can have as few as one independent element in the matrix ( $d_{11}$ ). The matrix takes its simplest form if the natural symmetry axes of the crystal are chosen for the coordinate system.

**Electromechanical coupling.** The direct piezoelectric effect makes a crystal a generator, and the converse effect makes it a motor. Consequently, a piezoelectric crystal has many properties in common with a motor-generator. For example, the electrical properties, such as the dielectric constant, depend upon the mechanical load; conversely, the mechanical properties, such as the elastic constants, depend upon the electric boundary conditions. The electromechanical coupling factor  $k$  can be defined as follows. Suppose electrodes are attached to a piezoelectric crystal and connected to a battery. Then the ratio of the energy stored in mechanical form to the electrical energy delivered by the battery is equal to  $k^2$ . In general,  $k$  ranges from below 1% to about 30%. In quartz, for example, the coupling is roughly 10%. In ferroelectric crystals,  $k$  can approach unity in certain circumstances. See FERROELECTRICS.

In quartz, a stress of 1 kg/cm<sup>2</sup> applied along the  $x$  axis produces a polarization of about  $2 \times 10^{-11}$  coulombs/cm<sup>2</sup> along the same axis. Conversely, an electric field of 100 volts/cm produces a strain of about  $2 \times 10^{-8}$ . In ferroelectric crystals, such as Rochelle salt and KH<sub>2</sub>PO<sub>4</sub>, and in certain antiferroelectrics, such as NH<sub>4</sub>H<sub>2</sub>PO<sub>4</sub> (ADP), these effects can be several orders of magnitude larger.

**Molecular theory.** Quantitative theories based on the detailed crystal structure are very involved. Qualitatively, however, the piezoelectric effect is readily understood for simple crystal structures. Figure 1 illustrates this for a particular cubic crystal, zinc blende (ZnS). Every Zn ion is positively

charged and is located in the center of a regular tetrahedron  $ABCD$ , the corners of which are the centers of sulfur ions, which are negatively charged. When this system is subjected to a shear stress in the  $xy$  plane, the edge  $AB$ , for example, is elongated, and the edge  $CD$  of the tetrahedron becomes shorter. Consequently, these edges are no longer equivalent, and the Zn ion will be displaced along the  $z$  axis, thus giving rise to an electric dipole moment. The dipole moments arising from different octahedra sum up because they all have the same orientation with respect to the axes  $xyz$ .

Another simple type of piezoelectric structure is encountered in barium titanate, BaTiO<sub>3</sub>, as shown in Fig. 2. The positive Ti ions are surrounded by an almost regular octahedron of negative oxygen ions. The Ti ions are not in the center of the octahedron, but somewhat displaced along the  $z$  axis. This structure already has a dipole moment or spontaneous polarization in the absence of externally applied stresses. It is clear from Fig. 2 that the Ti ion is pushed more off center when the crystal is mechanically compressed in the  $xy$  plane or elongated along  $z$ . The additional polarization associated with this deformation is the piezoelectric polarization. See BARIUM TITANATE.

**Piezoelectric ceramics.** Barium titanate and a few related compounds have the remarkable property that, by means of a sufficiently strong electric field, the direction of the spontaneous polarization can be switched to any one of the  $x$ ,  $y$ ,  $z$  axes. This makes it possible to produce polycrystalline samples (ceramics) which are piezoelectric. The electromechanical coupling factors of such ceramics can reach about 50%.

**Piezoelectric resonator.** The piezoelectric strains that can be induced by a static electric field are very small, except in certain ferroelectrics. Larger strains can be obtained when a piezoelectric crystal is driven by an alternating voltage, the frequency of which is equal to a mechanical resonance frequency of the crystal. The vibrating crystal reacts back on the circuit through the direct piezoelectric effect. In the range of a mechanical resonance, this reaction is equivalent to the response of the network shown in Fig. 3, provided that the series resonance frequency

$$f_R = 1/(2\pi\sqrt{LC})$$

of the network is equal to a mechanical resonance frequency of the crystal. An important difference between the network of Fig. 3 and the piezoelectric resonator is that the latter has many discrete modes of vibration, whereas the network has only one resonance frequency.

**Network elements.** The elements  $L$ ,  $C$ , and  $C_0$  of the equivalent network can be calculated from the physical constants of the crystal. Consider, for example, the simple resonator shown in Fig. 4. A rectangular crystal bar with the dimensions  $l_1 \gg l_2 \gg l_3$  is excited to compressional lengthwise vibrations. The  $xy$  faces have adherent electrodes, and the bar is oriented with respect to the

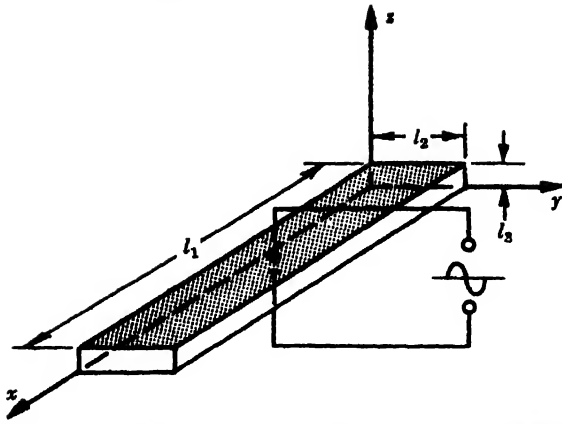


Fig. 4. Simple piezoelectric resonator. A voltage applied to the electrodes shortens or lengthens the bar, exciting longitudinal vibrations.

natural crystal axes so that an electric field  $E_3$  along  $z$  causes a strain  $S_1$  along the bar according to the equation  $S_{1(\text{piezoel})} = d_{31}E_3$ . A mechanical stress  $T_1$  along the bar causes a strain  $S_{1(\text{mech})} = s_{11}^E T_1$ , where  $s_{11}^E$  is the elastic compliance measured at constant electric field  $E_3$ . The resonance frequency for the fundamental lengthwise compressional mode is then

$$f_R = 1/(2l_1\sqrt{\rho s_{11}^E}) \text{ cps}$$

where  $\rho$  is the density of the crystal. The parallel capacitance  $C_0$  is the static capacitance of the crystal

$$C_0 = \frac{\epsilon l_1 l_2}{4\pi l_3 (9 \times 10^{11})} \text{ farad}$$

Here  $\epsilon$  is the relative dielectric constant along  $z$ . For  $C$  and  $L$ , the analysis yields

$$C = \frac{8d_{31}^2 l_1 l_2}{\pi^2 s_{11}^E l_3 (9 \times 10^{11})} \text{ farad}$$

$$L = \frac{\rho (s_{11}^E)^2 l_1 l_3 (9 \times 10^{11})}{8d_{31}^2 l_2} \text{ henry}$$

(All physical constants are in cgs units.) For the  $n$ th overtone,  $C_0$  and  $L$  remain the same, whereas  $C$  must be divided by  $n^2$ . The losses (damping) represented by the resistance  $R$  in Fig. 3 arise, for example, from ultrasonic radiation, friction in the crystal mount, internal friction in the crystal originating in various imperfections, and from dielectric relaxation.

At the mechanical resonance frequency  $f_R$ , the ac current is maximum and is determined by  $R$ . At the antiresonant frequency

$$f_A = \sqrt{(C_0 + C)/LC C_0}$$

the current is minimum. The difference  $\Delta f = f_A - f_R$  increases with increasing electromechanical coupling according to the equation

$$\Delta f \approx 4k^2/\pi^2$$

The reactance depends upon frequency, as

shown in Fig. 5. For a typical piezoelectric crystal, such as quartz, resonating at about  $10^6$  cps, the following orders of magnitude are typical for the elements of the equivalent network:

$$L \approx 10^2 \text{ henry}$$

$$C \approx 2 \times 10^{-14} \text{ farad}$$

$$C_0 \approx 5 \times 10^{-12} \text{ farad}$$

The damping resistance  $R$  varies from about  $10$  to  $10^4$  ohms; that is, the  $Q$  factors

$$Q = \frac{1}{R} \sqrt{\frac{L}{C}}$$

are in the range between  $10^6$  and  $10^4$ , and the resonances are very sharp. These characteristics can not be achieved with conventional coils and condensers as circuit elements.

**Vibration modes.** With piezoelectric resonators of various types, the range from audio frequencies to many megacycles per second can be covered. The vibration modes frequently used are (in order of increasing frequency) (1) flexural vibrations of bars and plates, (2) longitudinal vibrations of bars and plates, (3) face shear vibrations of plates, and (4) thickness shear vibrations and compressional vibrations of plates. Figure 6 illustrates some of these modes. The excitation of particular vibration modes can be achieved by proper orientation of the resonator with respect to the natural crystal axes, by proper positioning of the electrodes, and by proper mounting. A simple example is illustrated by Fig. 7. A bar is oriented so that an electric field along  $x$  causes an expansion or contraction along  $y$ . The electrodes are split and cross-connected so that the bar flexes in the  $yz$  plane when a voltage is applied. The fundamental flexure mode is easily excited with this arrangement. However, excitation of higher even-numbered flexural modes is also possible. Interesting resonators

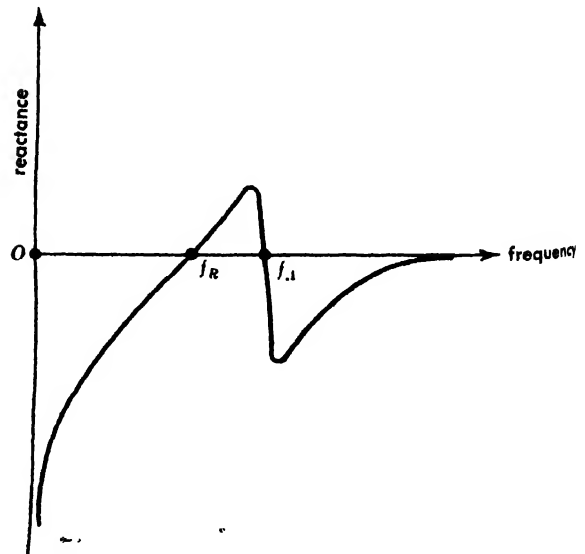


Fig. 5. Reactance vs. frequency for a piezoelectric resonator.



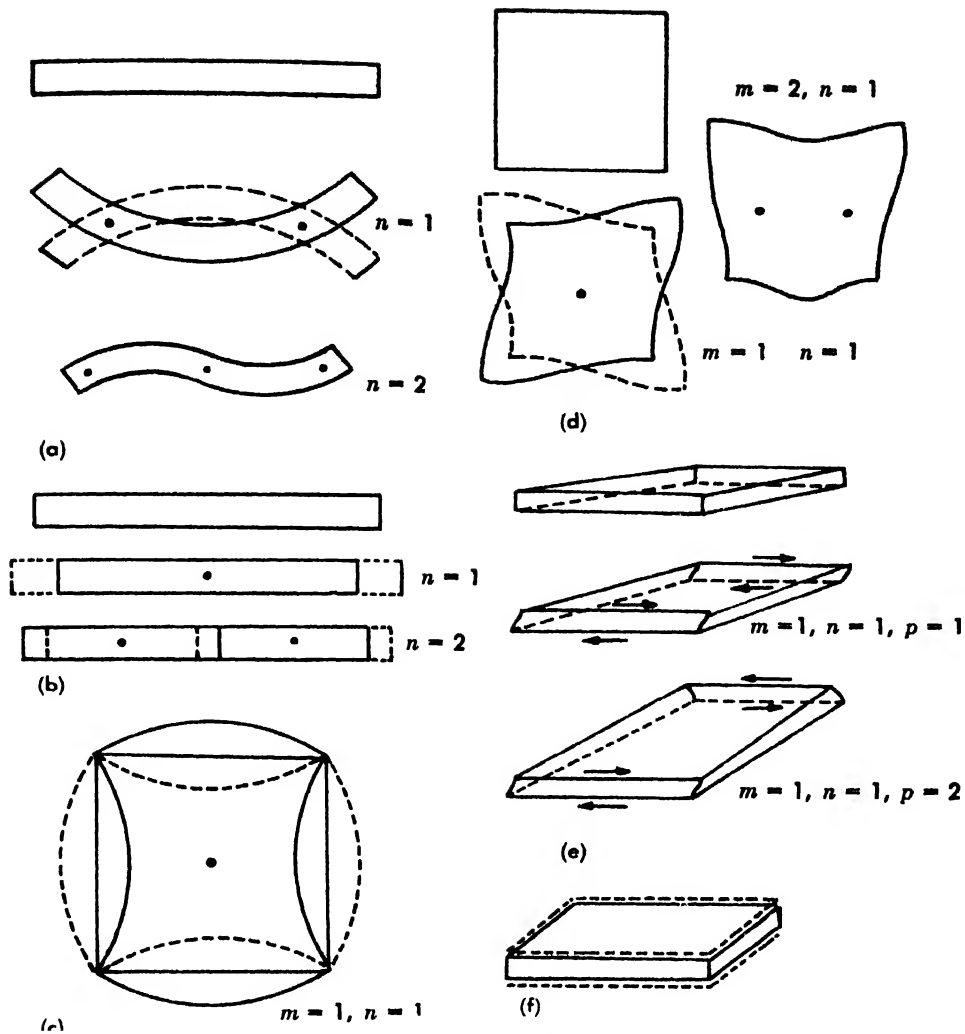


Fig. 6 Examples of vibration modes of bars and plates. (a) Flexural vibrations of a bar. (b) Longitudinal vibrations of a bar. (c) Longitudinal vibration of a

plate. (d) Face shear vibrations of a plate. (e) Thickness shear vibrations of a plate. (f) Thickness vibration of a plate.

are possible with piezoelectric ceramics ( $\text{BaTiO}_3$ -type) because different parts of the resonator can be polarized in different directions. See VIBRATION.

**Common applications.** The sharp resonance curve of a piezoelectric resonator makes it useful in the stabilization of the frequency of radio oscillators. Quartz crystals are used almost exclusively in this application. The main advantages of quartz are high  $Q$  factor, stability with respect to aging, and the possibility of orienting the resonator with respect to the natural crystal axes so that the temperature coefficient of the resonance frequency vanishes near the operating temperature. Figure 8 illustrates the orientation of commonly used cuts, and Fig. 9 shows the temperature dependence of the resonance frequency for a few of these.

In vacuum-tube oscillators, the crystal generally is part of the feedback circuit. In the circuit proposed by G. W. Pierce, the conditions for oscillation are not satisfied unless the crystal reactance is positive. Hence, the oscillation frequency is between the resonant and antiresonant frequency of

the crystal (see Fig. 5). Circuits of this type hold the frequency within a few parts per million. Much greater stability can be achieved with the bridge circuit of L. A. Meacham. Here the oscillation conditions are fulfilled for zero phase shift in the feedback circuit, that is, at the exact series resonance frequency of the crystal. Long-term frequency stability of about one part in  $10^8$  and short-term stability of one part in  $10^9$  can be achieved with such oscillators; for example, see QUARTZ CLOCK. For detailed information on the Pierce and Meacham circuits, see OSCILLATOR.

Selective band-pass filters with low losses can be built by using piezoelectric resonators as circuit elements. With a simple network consisting of resonating crystals only, a pass band of twice the difference between resonant and antiresonant frequency can be obtained. For quartz resonators, this pass band is about 0.8%. At relatively low operating frequencies, this band is too narrow, and combinations of crystal resonators with coils and condensers are generally used. A synthetic piezo-

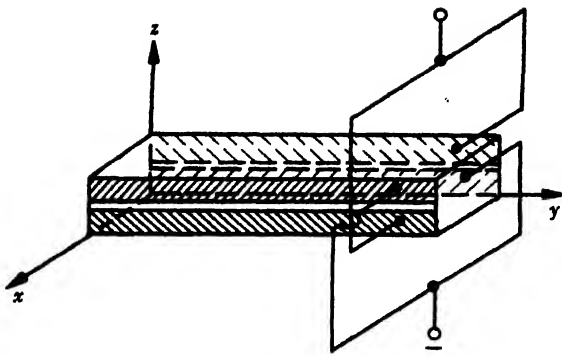


Fig. 7. Excitation of a flexure mode by means of split electrodes.

electric crystal which is often substituted for quartz in this particular application is ethylene diamine tartrate.

Piezoelectric crystals provide the most convenient means for generation and detection of vibrations in gases, liquids, and solids at frequencies above  $10^4$  cps. Quartz, ammonium dihydrogen phosphate, Rochelle salt, and barium titanate are frequently used in sonic and ultrasonic transducers. The mechanical impedances of liquids and solids are generally close enough to the mechanical impedance of the piezoelectric crystal so that efficient energy transfer is possible. The intensity of ultrasonic radiation that can be achieved is mainly limited by the mechanical strength of the piezo-

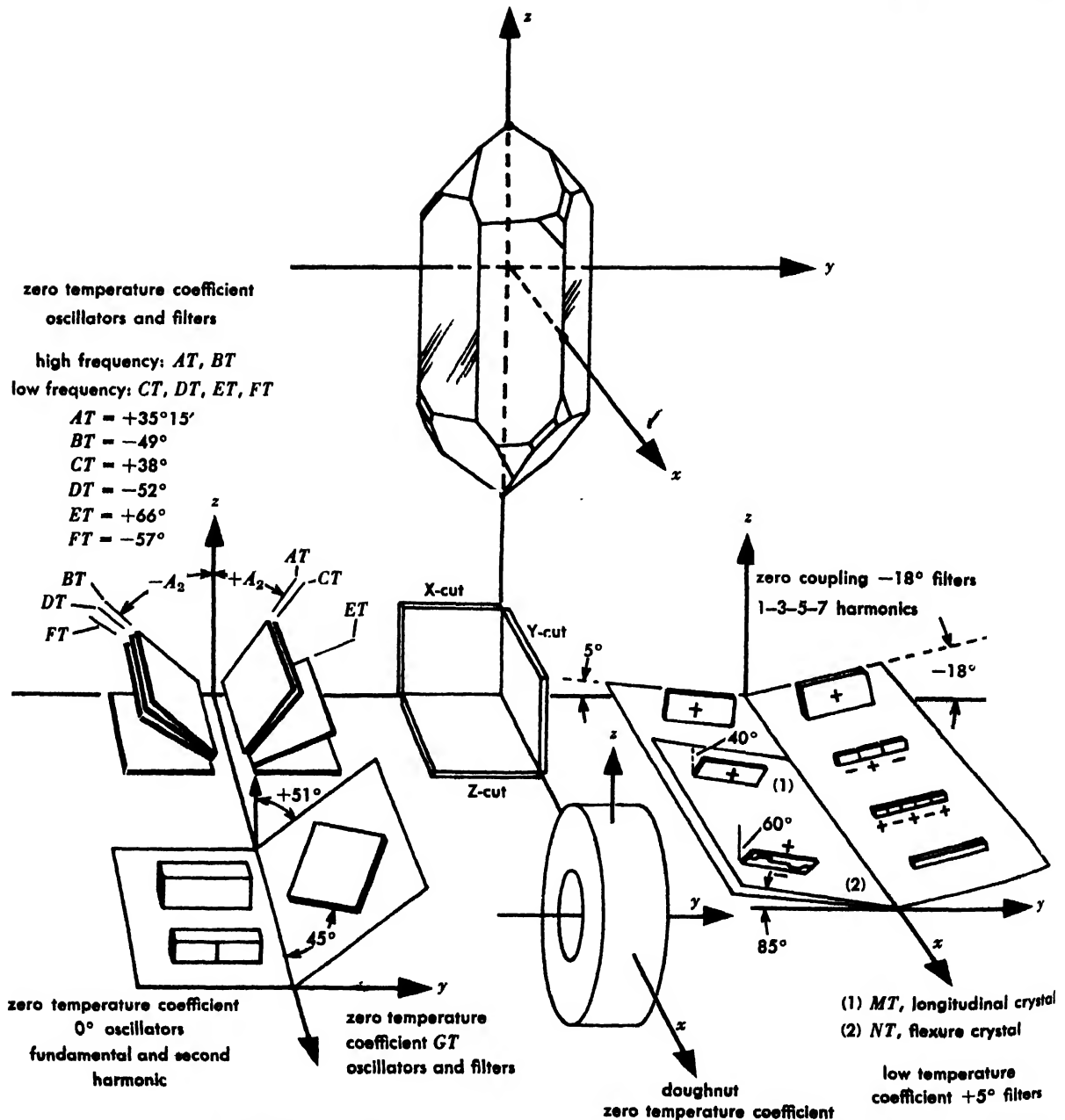


Fig. 8. Orientation with respect to the natural crystal axes of some of the more commonly used special cuts of quartz. (From W. P. Mason, *Piezoelectric Crystals*

and Their Application to Ultrasonics, Van Nostrand, 1950)

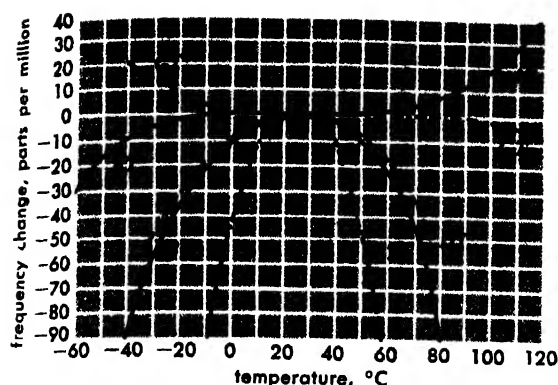


Fig. 9. Temperature dependence of the resonance frequency of quartz cuts commonly used in filters and oscillators. (From W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, Van Nostrand, 1950)

electric crystal. The maximum ultrasonic intensity theoretically obtainable in water by means of quartz or ammonium dihydrogen phosphate is of the order 2000 watts/cm<sup>2</sup> and 200 watts/cm<sup>2</sup>, respectively. For gases, the mechanical impedance match is so poor that the corresponding values are about 4000 times smaller. However, the mechanical impedance match can be greatly improved by using piezoelectric devices consisting of two differently oriented crystal cuts cemented together in such a way that a voltage applied to the electrodes causes the elements to deform in opposite directions, and a twisting or bending action results. Assemblies of this type (Bimorphs) with BaTiO<sub>3</sub> ceramics or Rochelle salt are widely used in such devices as microphones, earphones, and phonograph pickup cartridges.

Ultrasonic waves at microwave frequencies up to  $2.4 \times 10^{10}$  cps have been generated by means of the piezoelectric effect. The arrangement is shown in Fig. 10. The end surface of a piezoelectric crystal rod is exposed to a strong microwave electric

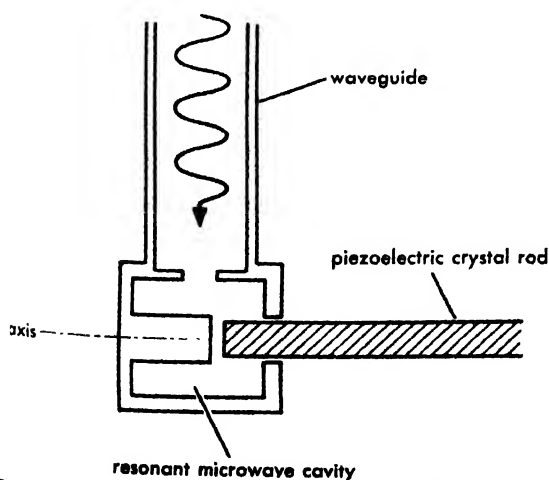


Fig. 10. Experimental arrangement for generation of ultrasound at microwave frequencies by means of a piezoelectric crystal.

field in a resonant reentrant cavity. The ultrasonic waves travel through the crystal rod in a guided wave mode. The attenuation is low only at very low temperatures. [W.K.]

**Bibliography:** W. G. Cady, *Piezoelectricity*, 1946; W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, 1950; J. F. Nye, *Physical Properties of Crystals: Their Representation by Tensors and Matrices*, 1957.

## Pigeonite

The name given to the monoclinic pyroxenes of the general formula (Mg,Fe)SiO<sub>3</sub> with some augite in solid solution. Pigeonite bears the same relation to the orthorhombic pyroxenes as augite does to the diopside-hedenbergite series. Pigeonite is the orthorhombic pyroxene equivalent in the volcanic rocks. Most high-temperature metamorphic and igneous orthorhombic pyroxenes were probably originally pigeonite. The small optic angle (2V) distinguishes the mineral from augite and the inclined extinction distinguishes it from the orthorhombic pyroxenes. Pigeonite forms black, brown, or dark green short stubby crystals with the 87° pyroxene (110) cleavages. The slower cooling rates of the igneous and metamorphic rocks usually permit the augitic materials in solution to exsolve and the remaining monoclinic pyroxene to invert to the orthorhombic form. The original augitic material is evident by the oriented exsolution lamellae in the host orthorhombic pyroxene. The faster cooling rates of the volcanic rocks quenches in the augitic material and thereby preserves the metastable pigeonite at surface temperatures. See ORTHORHOMBIC PYROXENE; PYROXENE; see also AUGITE; DIOPSIDE; ENSTATITE. [G.W.D.]

## Pigment

A finely divided material which contributes to the optical and other properties of surface coatings. Pigments are insoluble in the coating material, as opposed to dyes, which color the coating by dissolving in it. Pigments are mechanically mixed with the coating and deposited when the coating dries. They are not, in general, changed in their physical properties by incorporation in and deposition from the vehicle. Pigments may be classified according to their composition (inorganic or organic) or their source (natural or synthetic). However, the most useful classification is by their color as white, transparent, colored, or by their function.

**White pigments.** These materials are essentially transparent to visible light, but because of the difference in refractive index between the pigment particles and the vehicle, they refract the light from a multitude of surfaces and return a substantial portion in the direction of illumination without significant change in its spectral composition. See COLORIMETRY.

The common white pigments are titanium dioxide, derived from titanium-bearing ores; white lead from the corrosion of metallic lead; zinc oxide from the burning of zinc metal; and lithopone, a

mixture of zinc sulfide and barium sulfate. Less commonly, pure zinc sulfide and antimony oxide are used. Titanium dioxide may be crystallized in the rutile or the anatase form, depending upon the method of production. It may be modified further by surface treatment to control the rate of chalking and other properties. Rutile titanium dioxide has a higher refractive index than anatase, and therefore, higher hiding power, but it has a somewhat yellow color. Anatase gives a purer white.

White lead may be a basic carbonate of lead, made by corrosion of metallic lead by acetic or similar acids; or basic sulfate white lead, made by subliming lead ores under controlled conditions. The two pigments are largely interchangeable. Recently a basic silicate of lead has been made available, giving the same results with a somewhat lower lead content. White lead pigments are the oldest of the white pigments and are used in exterior paints, to which they contribute continued flexibility and durability.

Zinc oxide is commonly made by sublimation of metallic zinc in the presence of air. If a lead-containing ore is used, leaded zinc oxide is obtained, the lead being in the form of basic lead sulfate. Zinc oxides vary in particle size and shape, according to the intended use. Zinc oxide films are hard and brittle, and the pigment is rarely the sole pigment in a coating. A certain amount is usually used in exterior paints to contribute hardness. Zinc oxide was, for many years, the principal white pigment of interior paints, but has been largely superseded by titanium dioxide.

Lithopone is the pigment which results when solutions of barium sulfide and zinc sulfate are mixed, precipitating zinc sulfide and barium sulfate, both of which are water-insoluble. The pigment contains about 30% zinc sulfide, which is responsible for the hiding power. Although large amounts of this pigment were made and used in past years, it too has been largely superseded by titanium dioxide.

Pure zinc sulfide has a refractive index almost as high as that of titanium dioxide. It has, however, almost no advantages over titanium dioxide and is not widely used.

Antimony oxide is occasionally used in enamels but is most commonly found as an ingredient of certain fire-retardant paints.

**Transparent pigments.** The refractive indexes of these pigments are very close to the index of the paint vehicle (about 1.54). They are used to provide bulk, to control settling, and to contribute to durability, hardness, and abrasion resistance. They are often referred to as inert pigments, which is in error because some are extremely reactive. Because one of their common uses is to add bulk to other pigments, they are often called extenders. Most transparent pigments are natural minerals, reduced to pigment particle size. Among the most commonly used are calcium carbonate (ground limestone, whiting, or chalk), magnesium silicate, clay, silica, or barytes (barium sulfate). Barium carbonate and other minerals are sometimes used. Precipitated barium sulfate (blanc fixe) or calcium

carbonate are sometimes available, often as by-products of some chemical operation. Transparent pigments often constitute a substantial portion of a protective coating.

**Colored pigments.** These pigments are available in a wide variety of colors and properties, depending upon the end use. Several hundred have been used at one time or another, but the following are most common.

**Red.** Iron oxides, often classified by color, include Indian red, Spanish red, Persian Gulf red, and Venetian red (a mixture of iron oxide and calcium sulfate). Other red pigments include cadmium red (cadmium selenide) and organic reds, which are usually coal-tar derivatives, either precipitated in pigment form (toners) or deposited on a transparent pigment (lakes). Organic reds include toluidines and lithols.

**Orange.** Chrome orange (basic lead chromate) molybdate orange (lead chromate-molybdate), and various organic toners and lakes are the most common orange pigments.

**Brown.** Browns are nearly always iron oxides, although certain lakes and toners are used for special purposes.

**Yellow.** These pigments include natural iron oxides, such as ochre or sienna, or synthetic iron oxides, which are stronger and brighter, chrome yellow (normal lead chromate); cadmium yellow (cadmium sulfide); and organic toners and lakes such as Hansa Yellow and benzidine yellow.

**Green.** The most important green pigments are chrome green (a mixture of chrome yellow and Prussian blue); chromium oxide, which is duller but more permanent; phthalocyanine green, which is an organic pigment containing copper; and various other organic toners or lakes, often precipitated with phosphotungstic or phosphomolybdic acid.

**Blue.** The blue pigments include Prussian blue (ferric ferrocyanide, sometimes called milori or Chinese blue, depending upon the shade); ultramarine, an inorganic pigment made by fusing soda sulfur, and other materials under controlled conditions; phthalocyanine blue, an organic pigment containing copper; and numerous organic toners and lakes.

**Purples and violets.** These are nearly all organic toners or lakes. Manganese phosphate is an inorganic purple pigment but is very weak.

**Blacks.** The vast majority of black pigments consist of finely divided carbon (carbon black, lamp black, and bone black) usually obtained by allowing a smoky flame to impinge on a cold surface. Black iron oxide and certain organic pigments are used where special properties are required.

**Special pigments.** Anticorrosive pigments are used to prevent the formation or spread of rust on iron when the metal is exposed by a break in the coating. The most common are red lead, an oxide of lead, and zinc yellow or zinc chromate, a basic chromate of zinc. Other colored chromates are sometimes used. The color of red lead fades rapidly and the anticorrosive chromates are usually very

weak in tinting strength. Metallic lead is sometimes used for anticorrosive paint.

Metallic pigments are small, usually flat particles of metal, prepared for dispersal in coatings. Aluminum is most common, because it will leaf and form a smooth, metallic film. The flakes are sometimes colored. Bronze, copper, lead, nickel, stainless steel, and silver appear occasionally. Zinc dust, or powdered zinc, is used more often because of its excellent adhesion to galvanized iron than because of its appearance.

Luminous pigments will radiate visible light when exposed to ultraviolet light. Phosphorescent pigments continue to glow for some period after the exciting light has been removed; these are usually sulfides of zinc and other materials, with small amounts of additives which control the phosphorescent properties. Fluorescent pigments lose their luminosity as soon as the exciting light is removed; these may be sulfides, also, but many organic pigments have this property. See LUMINOUS PAINT.

There are a number of other specialized pigments, such as those which change color at some predetermined temperature for use in indicating hot areas on motors, pigments to give a pearly appearance, and pigments which conduct electricity for printed circuits.

Coarse materials, such as pumice, are often added when a non-slippery coating is required. Glass beads give a very high degree of refractivity in the direction of illumination and are often used in center-line paints for signs where night visibility is required. Intumescent pigments puff up under heat, giving a fire-resistant coating. See COLOR; DYE; PAINT; SURFACE COATING. [F. S. D.]

*Bibliography:* W. Von Fischer (ed.), *Paint and Varnish Technology*, 1948.

## Pigmentation

The normal color of the body and its organs results from a summation of the natural color of the tissue, the pigments deposited therein, and the pigments carried through in the blood bathing it. Of the pigments present in the skin and hair, melanin is most important, while hemoglobin in its reduced and oxidized form is the most important blood-borne pigment. See CHROMATOPHORE.

**Melanin.** This is a pigment present as microscopic or submicroscopic brown or black granules in the cells of the superficial layers of the skin, occasionally in the underlying dermis, in the hair, and in the iris. It appears to be localized in the mitochondria and is a polymer of cyclic organic compounds derived from the amino acid tyrosine by the action of the enzyme tyrosinase. The effect of melanin on skin color depends not only on its own inherent color but also on its ability to scatter incident light. The bright colors of the perineum of baboons, for example, are a result of the scattering of light from dark pigment granules in specific spatial arrangements.

The amount of melanin in the skin varies within wide limits from species to species, from race to race, within members of a given stock, and from

site to site on the body, in addition to being subject to change when stimulated by a variety of forces. The dark color of the negroid skin is due to the concentration of melanin in the skin; the dark color of the eyelids, nipples, and genitalia in Caucasians has exactly the same cause. Exposure to ultraviolet rays, as in sunlight, or to other ionizing radiation causes increase in the amount of melanin present in the skin.

The color of the hair depends on the relationship of the amount and concentration of pigment and air: the darker the hair, the more pigment and less air. The gray or white hair of old age contains almost no melanin.

**Control of pigmentation.** In many animals, probably including mammals, pigmentation is at least to some extent controlled by hormones, notably a proteinaceous hormone of the pituitary gland known as melanocyte-stimulating hormone (MSH). This hormone (or similar hormones), which has a striking effect on skin color in amphibians and fish, presumably plays a part in the generalized changes in pigmentation which accompany some conditions in man caused by hormonal imbalance. In these conditions there may be a direct or indirect influence on MSH production. Other conditions associated with variations in the steroid sex hormones and of the thyroid hormone are also associated with changes in skin pigmentation.

Another possible controlling force over skin pigmentation is the nervous system. The cells which produce melanin in the body have the same embryological derivation as nervous tissue, and indirect evidence suggests that there is some neural control over these cells.

The basic pattern of pigmentation is, of course, hereditarily determined, and some abnormalities of pigmentation have a similar hereditary basis.

**Focal abnormalities.** In melanin pigmentation, focal abnormalities include the familiar freckles and the dark skin blemishes known as moles or nevi. These lesions reflect focal excessive deposition of melanin and they are entirely benign, but a malignant counterpart exists in the malignant melanoma which appears to come about as wild uncontrolled overgrowth of the melanin-producing cells. Focal spots of hyperpigmentation also occur commonly in response to inflammation or irritation. Among other focal abnormalities are the irregular patches of depigmentation, occasionally present from birth (leukoderma), or arising later in life as a result of infection, trauma, or some unknown causes (vitiligo). Spots of hyperpigmentation of varying sizes and shapes occur in conjunction with such varied diseases as polyostotic fibrous dysplasia (a disease of bone with an associated endocrine abnormality), neurofibromatosis (multiple widespread benign tumors of neural origin), intestinal polyposis (a heritable condition characterized by multiple polyps of the intestine), and in association with various internal malignant tumors.

**Generalized abnormalities.** In external melanin pigmentation affecting the entire body surface, generalized abnormalities are not uncommon. The

complete or partial absence of melanin found in albinos is an inborn disorder inherited as a Mendelian recessive or, in mild cases, as an irregular dominant caused by the lack of the enzyme tyrosinase which is necessary to convert tyrosine to melanin. The lack of pigmentation in the eyes of these subjects is commonly associated with ocular difficulties. In phenylpyruvic oligophrenia, a rare recessive inborn disorder characterized by mental deficiency, there is a lack of tyrosine because the enzyme necessary to convert the amino acid phenylalanine to tyrosine is lacking and, although tyrosinase is not diminished, the usual amounts of melanin cannot be formed. The patient is therefore pale.

Generalized hyperpigmentation occurs in adrenal hypofunction (Addison's disease), probably indirectly through an effect on MSH, while hypopituitarism causes hypopigmentation because MSH production is suppressed in the underactive pituitary gland. The changes in pigmentation of the nipple and other areas in pregnancy, as well as pigmentation disturbances in eunuchs, point to a control, either direct or indirect, of melanin production by the steroid hormones. Melanin deposition in the skin is also increased in hemochromatosis, although part of the pigmentation in this disease is due to the deposition of iron-containing blood pigments. Certain intoxications, notably arsenic poisoning, can also lead to generalized hyperpigmentation.

**Blood-borne pigments.** Of the pigments normally present in the blood stream, hemoglobin, the oxygen-carrying pigment of the red blood cells, is the most important in determining skin color. Its effect is best appreciated in the opposite phenomena of blushing and blanching, in which an increased or decreased volume of blood and therefore of hemoglobin, perfuses the skin because of dilation or constriction of the skin vessels.

The red portions of the body, the mucous membranes, lips and similar structures, and the red organs, owe their color to the propinquity of blood-carrying vessels to the surface. Similar pigmentation is seen in hemangiomas, the so-called strawberry marks and similar skin blemishes, in which a local abnormal proliferation of blood vessels occurs close to the skin surface.

Qualitative and quantitative changes in circulating hemoglobin are reflected in visible changes in skin color. In anemia, for instance, when there is a decrease in this pigment, the body is pale; while in polycythemia, a condition characterized by increase in circulating red cells, the skin is ruddy. Oxidized hemoglobin, that is, hemoglobin carrying oxygen from the lungs to the body tissues, is bright red, while reduced hemoglobin is purplish. In conditions in which the blood is poorly oxygenated because of disease of the lungs or heart or through the inability of the blood to circulate to the lungs (so-called blue babies), the skin becomes purple or blue (cyanosis). When carbon monoxide is present in inspired air, it combines with hemoglobin to

yield a cherry-red pigment, carboxyhemoglobin and when certain reducing substances are introduced into the blood stream, another abnormal hemoglobin product, methemoglobin, which has chocolate color, is formed. The presence of these substances in the blood is reflected in appropriate changes in skin color.

Other pigments may be carried in the blood stream. Jaundice, a yellowing of the skin, is caused by the deposition of bile pigments in the skin. Carotenemia, another form of yellowing of the skin is found in individuals consuming large quantities of carotene-containing fruits and vegetables such as carrots and apricots. Small quantities of carotene are normally present in skin and contribute slightly yellow color to it. See JAUNDICE.

**Exogenous pigments.** Finally, skin color may be changed by the deposition of entirely foreign substances which may be introduced directly, as in tattoos, or indirectly through ingestion, injection or inhalation, as in silver poisoning (argyria), lead poisoning (plumbism), and bismuth poisoning, in which the metallic salts themselves are deposited in the skin to cause pigmentation.

**Other organs.** While other organs besides the skin may undergo color change, the color changes are analogous to those seen in the skin. The basic color of an organ is determined by a combination of its fat content (yellow), blood (red or purple) and native pigments (ceroid, lipochrome, cardiolipin, carotene, hemosiderin), and by the scattering of light from microscopic and submicroscopic particles. The organs may change color because of increases or decreases in any of these parameters through introduction of foreign substances, as in pigments (anthracosis of the lungs) and products of catabolism (melanosis of the colon) or through accumulation of colored material normally present in minute amounts (alkaptonuria or ochronosis). [R.B.H.]

## Pile foundation

A substructure supported on structural units introduced into the ground to transmit a load to lower strata or to alter the physical properties of the ground (Fig. 1). The supporting units, called piles, are made of wood, concrete, or steel.

**End-bearing and friction piles.** End-bearing piles transfer load from the footing to bedrock or other firm material capable of withstanding the intensity of pressure transmitted by the pile. Certain types of cast-in-place concrete piles can be formed with an enlarged base which will reduce the intensity of applied pressure and thus permit the use of end bearing of material of less strength or allow the use of a greater applied load.

Friction piles are driven into fairly deep beds of soil of somewhat uniform consistency. They transfer their load to the surrounding ground by means of friction along the length of embedment or by shearing of the closely surrounding soil. The full length of the pile may be effective in friction or the pile may extend above the level of the friction.



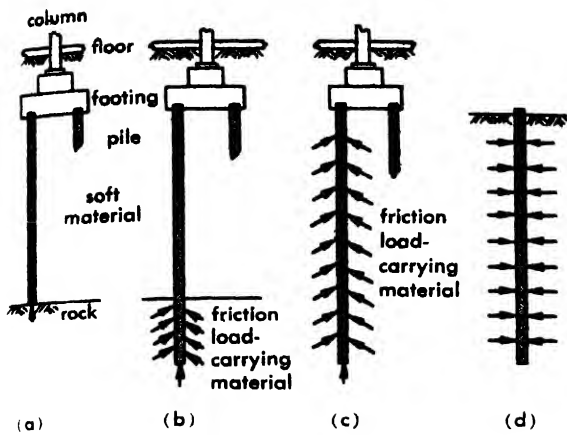


Fig. 1. Uses of piles. (a) As end-bearing column. (b) As friction pile in lower portion. (c) As friction pile full length. (d) As compactor of soil.

tional material through unstable strata, water, or air. Piles may act as compactors of some types of soil, thus improving its sustaining power.

Bearing value of a pile is the allowable or presumptive load which a particular type, length, and size of pile, at a given spacing, can sustain in the specific ground conditions at a site without the settlement exceeding an acceptable amount. It depends upon the ability of the ground to receive the load from the pile in friction or end bearing, and upon the ability of the soil to distribute the load until available resistance is mobilized within an acceptable amount of settlement. The soil is as much a part of a foundation as is the pile, and soil properties must be considered. *See SOIL MECHANICS.*

All piles distribute load to the soil in the shape of a so-called bulb of pressure, which may be shown by lines connecting points of equal stress intensities (Fig. 2). Where pile spacings are such that bulbs for individual piles overlap, they merge into a larger bulb extending to a greater depth. Overlaps also result in larger intensities of pressure at certain levels. This tends to produce greater settlement. Settlement produced by a uniform load increases in proportion to the diameter of the loaded area for cohesive soils, but the size has little effect in cohesionless soils. In cohesive soils, settlements continue over long periods of time, and because they are dependent upon intensities of pressure, vary with load and location under the structure. The allowable differential settlements should be established in cohesive soils.

Bearing capacities are usually computed by formulas, the most common of which assumes the ultimate bearing value to be equal to the dynamic driving force, using factors for energy losses in hammer, pile, and impact. Such formulas are applicable only with cohesionless soils. Bearing capacities may be estimated by judgment as to unit friction values based upon hardness of driving the sampler when making soil test borings, or by laboratory tests of soil samples. Most accurate results are obtained from field loading tests on an actual

pile, from which the allowable load is determined by dividing the load at failure by a factor of safety. Load tests on cohesive soils, however, only test the shearing value of the soil or pile friction, and give no indication of the ultimate settlements.

**Pile driving.** This is the driving of piles, or casings or shells to be filled later with concrete, by impact from a pile-driving hammer. The term is sometimes loosely applied to the placing of piles by any method.

Drop hammers consist of a weight raised by a cable running over a frame and extending back to a drum. The weight is released by a trip or the drum is allowed to unwind rapidly. Drop distances are usually high. Frequency of blows is low.

Steam hammers may be of several types, all of which require the use of steam boilers or air compressors. Single-acting hammers employ steam or compressed air to raise the movable mass which is then tripped and falls by gravity alone. In the United States, the movable ram is inside the casing; in Europe, the casing is sometimes the movable part. Characteristics of the blow are a low striking velocity by a heavy weight. Double-acting hammers use steam or compressed air to raise the striking part and also to impart additional energy during the downstroke. They run at greater speeds than do single-acting hammers. Differential-acting hammers also employ steam or air to raise the striking part and impart additional energy on the downstroke but avoid a drop from the entering steam pressure to mean effective steam pressure.

Diesel hammers are self-contained units, made up of cylinder, piston or ram, small fuel tank, and fuel injector. They weigh less than steam hammers and require no steam or air supply. They use small amounts of fuel oil.

Jetting consists of displacing the soil at the pile tip by means of discharge of a quantity of water or air through an internal or external jet pipe, which also lubricates the sides of the pile as it rises to escape. It is done to aid the pile in passing through hard strata not situated suitably to be bearing material or to obtain embedment in hard material which might cause damage to the pile.

Jacking piles down requires a resistance to thrust against. This method is used in underpinning, where short sections of pile are inserted and

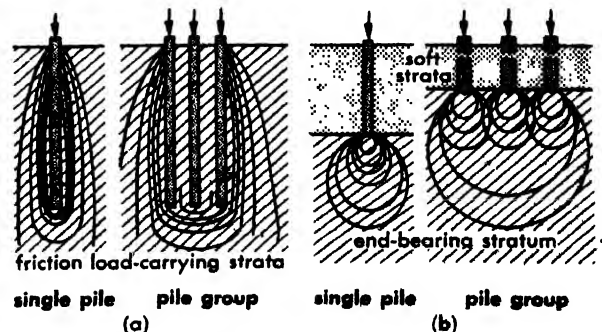


Fig. 2. Bulbs of pressure. (a) For friction piles. (b) For end-bearing piles.



connected after each is jacked down. The resistance of the pile is measured by a gage connected to the jack. This method avoids vibration from driving.

Drilling is used to install piles. A drilled pile is a cast-in-place pile constructed by pouring concrete in a drilled shaft. This avoids vibration or displacement and heave of the ground. Noise from driving is reduced. Holes can be drilled where restricted headroom would prevent driving long piles. A bell-shaped cut at the bottom increases bearing area and pile capacity. Drilling equipment is light and easily portable, or is truck-mounted. Precast piles are sometimes set in drilled or cored holes and grouted in place.

**Types of piles.** Wood piles, cut from straight tree trunks, are widely used. Most commonly used in North America are southern pine, Douglas-fir, oak, and cypress. Wood piles buried in the ground are subject to decay and insect attack unless submerged. Wood piles above ground are also subject to fire. In marine locations, piles are exposed to decay, abrasion, and marine borer attack. As a preservative, wood piles are often treated with creosote or a creosote-coal tar mixture applied under pressure. Mechanical protection in the form of concrete jacketing, pipe sleeves, or metal sheathing is also sometimes used on marine piles. Allowable loads on wood piles are usually less than those on concrete or steel piles, and they will not stand as hard driving.

Precast concrete piles are square, octagonal, or round. They may be of uniform cross section, or tapered full length, or may have tapered points. Prestressed concrete piles may be used. Sectional hollow precast concrete piles are made of hollow-spun reinforced concrete sections strung together on tensioned steel cables running through longitudinal holes.

Cast-in-place concrete piles need no storage space or special handling equipment, avoid the cutting off or extending involved in making length adjustments, and cannot be damaged by driving. Uncased piles may be used where it is certain that neither soil nor water will fall into the hole or squeeze into it and reduce the size. One method is to drive a casing pipe and core, then pull the core, place concrete, set the core on the concrete, and pull the casing; a concrete pedestal may be driven out if desired. Cased piles are used where support is required for the sides of the hole and where water occurs. A temporary heavy outer casing is driven and a permanent light shell inserted. In one method, a core is driven with the temporary casing to displace the soil; in other methods, concrete or metal bases are used to keep soil out of the casing. Thin metal shells may be driven by internal mandrels which are withdrawn and the shells filled with concrete.

Pipe piles consist of steel pipe of 6–30 in. in diameter. Ends may be closed by shoes or open. Open-type piles are usually cleared of soil by jetting. Both types are usually filled with concrete. Pipe piles are capable of carrying large loads to

rock, and lengths of 200 ft have been used. They are used where rock is within reach for end bearing, where water conditions would require use of air if caissons were sunk, and where soil displacement must be avoided. They are often used for underpinning, where they have been driven in sections as short as 2 ft in restricted headroom.

**Pile designs.** H piles are rolled steel section frequently used as piles. They are suitable for penetrating rock or other hard material with less effort. There is a wide range of sizes and weights. Lengths have exceeded 200 ft. Protection against corrosion in upper portions may be required.

Box piles are made by welding together two sections of steel sheet piling, or combinations of beams, channels, and plates. They perform the same functions as pipe piles. They are usually not filled with concrete but can be cleaned out and filled to any depth for strength and protection of the interior against corrosion.

Composite piles are piles in which the upper and lower portions consist of different types of piles. This is done to secure the particular advantages of a certain type without the expense of using it throughout. For example, concrete upper sections may be used on wood lower sections, where the latter would decay if they were near the surface. H or pipe lower sections may be used under cased concrete piles where it would be too expensive to use concrete alone.

Sheet piling is a vertical wall composed of interlocking units used to retain a difference in ground elevation, in which the bottom support is obtained by driving the toe below the lower level. Sheet piling may be wood, steel, or concrete, and may be temporary or permanent. Factors affecting choice include cost, ease of installation, availability, salvage, corrosion, decay, ability to withstand driving, lateral strength, and ease of making connections. Sheet piling is driven by pile hammers. Alignment and resistance to thrusts are provided by horizontal walers and braces or tiebacks.

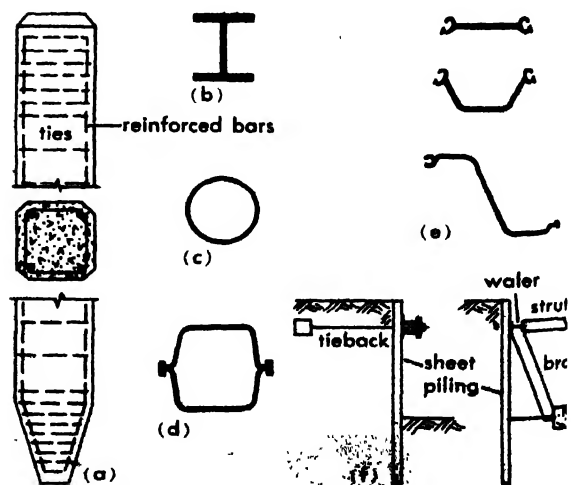


Fig. 3. Pile design. (a) Precast concrete pile. (b) H pile. (c) Pipe pile. (d) Box pile. (e) Steel sheet piling sections. (f) Sheet piling walls.

Wood sheet piling consists of one, two, or three thicknesses of plank with various types of edge connections. It is largely used for building foundation excavations and cofferdams.

Steel sheet piling is formed of flat, channel- or Z-shaped sections provided with edge interlock connections, having a wide range of strength. It is used for cofferdams, piers, retaining walls, caissons and excavations.

Concrete sheet piling consists of precast, rectangular, concrete piles with edge interlocks. It is often used in shoreline work. See CAISSON FOUNDATION, COFFERDAM; RETAINING WALL. [R.D.C.]

*Bibliography:* See FOUNDATIONS.

## Pilot production

In mass-production industries where complicated products, processes, or equipment are being developed a pilot plant often leads to the presentation of a better product to the customer, lower development and manufacturing costs, more efficient factory operations, and earlier introduction of the product. Following the engineering development of a product, process, or complicated piece of equipment, and its one-of-a-kind fabrication in the model shop, it becomes desirable and necessary to "prove out" the development on a simulated factory basis.

**Facilities.** To accomplish this, the pilot plant, an intermediate step between development laboratory and production factory, is established. It is provided with personnel and facilities that duplicate as nearly as possible actual manufacturing conditions but are free of the day-to-day necessity of meeting delivery schedules.

This requires personnel who are highly skilled in their field, but are not necessarily experts. They must understand the thinking and objectives of development engineers, and must also appreciate the day-to-day problems that arise in factory operation with unskilled or semiskilled operators.

Adequate equipment and services should be provided to accommodate the full range of factory requirements. Sometimes, however, it may be impractical to include certain facilities in a pilot plant because of floor space limitations, or because of specialized services required. In such cases, arrangements should be made to perform these pilot plant functions on regular production equipment in the factory, but under pilot plant supervision and financial control.

**Objectives.** Some of the objectives of a pilot plant operation are as follows.

**Quality control.** Pilot plant operation demonstrates the ability of processes or equipment to function consistently at the desired quality level under factory operating conditions.

**Material usage.** To demonstrate economical usage of materials during the pilot operation, the materials used should be truly representative and contain the full range of variability permitted by material purchase specifications.

**Process reliability.** Pilot operation demonstrates whether the processes involved are practical and

realistic and whether operating procedures can be followed in their entirety by the type of personnel used in the factory.

**Equipment reliability.** To demonstrate the capability of the equipment to produce at speeds and machine efficiencies, which will become standard in the factory, data are recorded during pilot production which show whether the equipment will perform at a satisfactory rate of defectives. Pilot plant experience establishes a list of spare parts which will be initially provided in the factory.

**Personnel requirements.** Pilot experience establishes the amount and labor grade of production and maintenance personnel required to operate the process or equipment. This is done with the cooperation of industrial engineering and industrial relations departments.

**Safety.** During pilot runs the safety department has an opportunity to review the process or equipment for compliance with safety requirements.

**Factory acceptance.** Perhaps the most important objective of a pilot plant is to afford representatives of the factory to which the process or equipment will be transferred an opportunity to witness the operation and to agree that it is ready for the factory. Thus, criticism of the development is minimized and its introduction into regular factory operation accomplished more quickly.

**Manufacturing costs.** Pilot plant operation provides the opportunity to verify engineering estimates of manufacturing cost and may indicate the necessity for pricing changes or for taking further steps to reduce the cost of product in order to reach a satisfactory level of profit.

**Development costs.** The cost of designing and developing processes or equipment and of instructing personnel in their use is often only a fraction of the total cost involved in producing a final efficient process or machine. The debugging or working out of unforeseen problems can consume a great deal of time and add a large amount of expense before the process or machine is ready for factory operation. The pilot plant relieves the factory of the expense of a trial run both in direct cost and in lost production of established products. The close proximity of the pilot operation to engineering and other skilled personnel permits first-hand observation of operation and quicker decisions on required changes. It may also save the expense and time of such personnel traveling to factory locations and the return of equipment from the factory to its point of origin.

The financing of a pilot plant should be on a budgetary basis. Funds for its facilities, personnel, and operation should be provided as part of a manufacturing development budget. The pilot plant should not be expected to finance itself through manufactured products, otherwise management tends to keep the project in the pilot plant longer than necessary in order to create funds for its operation. The pilot plant then loses its value as a development function, and the desire and interest of its personnel to introduce improvements and to tackle new projects becomes subordinate to that of

meeting production schedules. On the other hand, arrangements should be made to credit the operation with a fair value of the product manufactured so that the net amount remaining will represent development cost more accurately. See PRODUCTION ENGINEERING. [J.E.W.]

## Piloting

A form of navigation by which position is determined relative to external reference points, usually fixed points on the earth. It is the oldest form of navigation. Primitive man learned to guide himself by means of prominent rocks, trees, bends in the river, and other visible landmarks before he had any instruments other than his five senses. Piloting, usually called pilotage by aviators, remained little changed until the application of electronics to navigation in relatively recent years. Before that time, piloting was associated with nearness to land or other dangers to navigation, and frequent or continuous determination of position or positional information.

Position of a craft is determined in three ways: dead reckoning, celestial navigation, and piloting. Dead reckoning is the advancement of a previous position for subsequent motion of the craft. Celestial navigation uses celestial bodies, determination of position being with respect to the terrestrial points having the bodies momentarily overhead. Piloting uses points, generally fixed points on or near the surface of the earth, for determination of position. Although modern navigational aids and electronic devices have altered techniques somewhat, navigation may still be classified according to these three traditional methods.

Piloting and celestial navigation are different despite some similarities. In piloting, position is determined relative to the points upon which certain measurements are made (sometimes indirectly); whereas in celestial navigation, the objects observed merely provide certain information which is used to establish position relative to other points. This distinction applies even in space navigation.

**Points or marks.** Natural landmarks, the only piloting aids available to primitive man, have been supplemented with a great many man-made aids to navigation. Such aids are of several types. Beacons, both lighted and unlighted, and lighthouses are at fixed positions on land or in the water. Buoys, both lighted and unlighted, and lightships float in the water, being moored at desired points. Unlighted aids are called daymarks. These are given distinctive shapes or coloring to assist in identification. A light installed as an aid to navigation is given a distinctive sequence and duration of light and dark periods and color or colors, by which it can be identified. These characteristics are shown on the charts and given in the light lists available to navigators.

In addition to visible aids to navigation, bottom topography can be of assistance in locating the position of a vessel. Aircraft aids to navigation consist principally of electronic aids and lighted

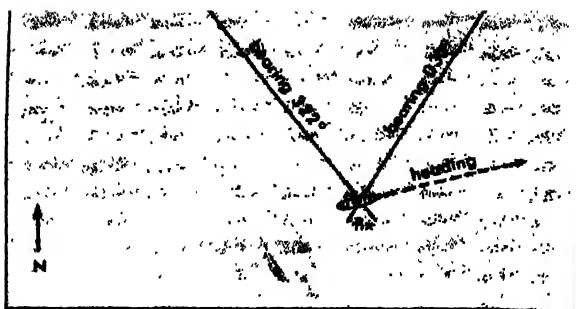


Fig. 1. A fix from two bearings.

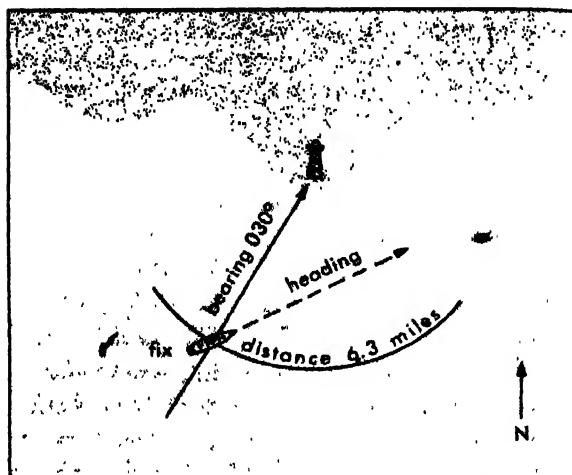


Fig. 2. A fix from a bearing and distance of one object.

beacons. Electronic aids are chiefly beacons of various kinds, airway radio ranges, and the various electronic systems of navigation. See NAVIGATION SYSTEMS, ELECTRONIC.

**Measurements.** The measurements made for piloting purposes are of direction, distance, differential distance between two points, and distance to the bottom aboard ship or to the surface of the earth in aircraft. In some electronic systems these measurements are made indirectly, as by matching pulses on an electroscope, counting dots and dashes, or centering pointers.

**Direction.** The direction of a landmark is called its bearing, commonly stated as the angular distance from some reference direction, usually true, magnetic, compass, or gyro north. Bearing values are generally given in integral degrees (although half and quarter degrees may be used), using three figures, from 000° at north clockwise through 360°. Relative bearings use the heading of the craft as the reference direction, and may be stated as above, right and left through 180°, or with reference to some part of the craft (as

on the starboard bow, 10 degrees abaft the port beam, or two points on the port quarter).

Aboard ship, bearings are usually measured (1) by noting when two objects are in range (directly in line); (2) by means of a suitable attachment to a compass or compass repeater; (3) by pelorus, a compasslike instrument without directive properties; or (4) electronically by radio direction finder, radar, or by the indication of the receiver-indicator of an electronic system of navigation.

**Distance.** This is now generally measured by radar in nautical miles or yards (meters in countries using the metric system). See RADAR.

**Differential distance.** Between two points, such distance is usually measured electronically by means of the receiver-indicator of a hyperbolic navigation system. The reading is generally in some special unit such as a microsecond (one millionth of a second) or one hundredth of a lane width. See HYPERBOLIC NAVIGATION SYSTEM.

**Marine depth.** Depth of water is measured in feet or fathoms (6 ft). Meters or special units are used by mariners of some countries. Measurement may be made by hand lead (pronounced lēd), a lead weight attached to a line; sounding machine, a reel of wire to one end of which is attached a weight which carries a device for recording depth; or, more commonly by echo sounder, a device which emits a sonic or ultrasonic signal in the water and times the return of the echo. See ECHO SOUNDER.

**Air altitude.** Height of an aircraft is measured in feet (meters by navigators of some countries), by means of an altimeter. A barometric altimeter determines height above sea level or some other reference level by measuring the atmospheric pressure at the aircraft. An absolute altimeter determines height above the terrain, usually by measuring the time interval between transmission of a signal and the return of its echo from the surface, or by measuring the phase difference between the transmitted signal and its echo. See ALTIMETER, PRESSURE; ALTIMETER, RADIO.

**Position determination.** In marine piloting, position is generally determined by means of lines of position, each indicating a series of possible positions of the craft at the time of measurement. A measured bearing provides a straight line of position (actually part of a great circle) passing through the object sighted. A measured distance provides a circular line of position with the object as the center and the distance as the radius. A measured differential distance provides a hyperbolic line of position. While a straight or circular line of position is usually plotted directly on the nautical chart, a special chart or table is generally used to interpret a hyperbolic line of position. Depth of water or height of an aircraft generally does not provide a line of position unless the craft crosses a distinctive feature such as the 100-fathom curve or a mountain ridge.

A position, called a fix, is determined by crossing two or more lines of position taken simultane-

ously or nearly so. If one line is adjusted for an appreciable time, a running fix is obtained. This is somewhat less reliable than a fix. A two-bearing fix is shown in Fig. 1. Figure 2 shows a fix obtained by means of a bearing and distance of a single object.

In aircraft, piloting is performed somewhat differently. When visual landmarks are used, they are generally checked off successively from a list or aeronautical chart as they are passed. Radar and radio marker beacons are used in much the same manner. Hyperbolic and other long-distance systems are used as in marine navigation. Modern radio ranges and some landing aids are used by keeping the needles of special indicators centered. Visual and aural signals are used in some older ranges. Airborne radio direction finders are generally of the automatic type providing a continuous indication of direction of the transmitter. See DIRECTION-FINDING EQUIPMENT.

Positions, either aboard ship or in the air, may be obtained in other ways than by crossing lines of position. Certain combinations of successive bearings of the same object provide immediate indication of position. Simultaneous horizontal angles between three objects can be set on a three-arm protractor (called a station pointer by British navigators) and the position determined by fitting the arms of the instrument to the chart symbols representing the objects. The horizontal angles are usually taken by marine sextant. This provides an accurate position sometimes used in hydrographic surveying or in locating the point of letting go an anchor, but seldom for other purposes. A line of soundings plotted on a transparency at the correct distance intervals at the scale of the chart and then matched by trial and error to the soundings shown on the chart may provide a reliable indication of position where the bottom topography has a distinctive pattern. Under favorable conditions, a bottom profile obtained by a recording echo sounder can provide a better position.

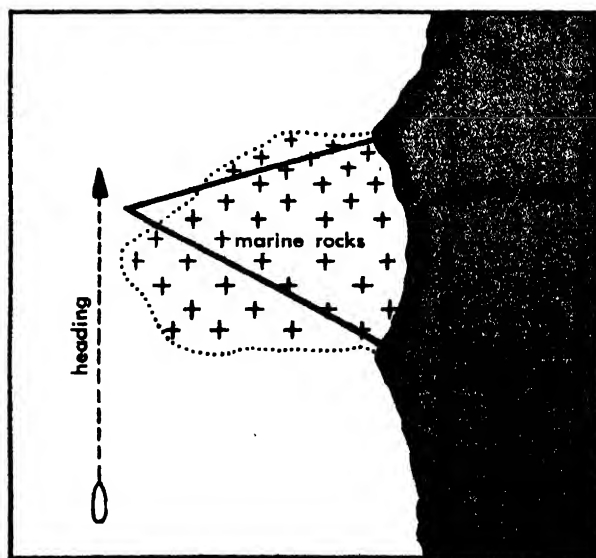


Fig. 3. A horizontal danger angle.

A vessel can sometimes be kept in safe water without a fix. In an area where the bottom shoals gradually, a well-chosen danger sounding can give warning in sufficient time to prevent grounding. A danger bearing of some object well ahead can provide an indication of a dangerous situation if the measured bearing exceeds the danger bearing. An off-lying rock or shoal can sometimes be successfully avoided by keeping the measured angle between two objects less (or more) than a preselected horizontal danger angle, as illustrated in Fig. 3. Similarly, a vertical danger angle can be established between the top and bottom of a single object.

No other form of navigation presents the variety of methods, or taxes the ingenuity of the navigator, as does piloting. In no other form is judgment, based upon experience, so valuable. See BUOY; CELESTIAL NAVIGATION; DEAD RECKONING; LIGHTHOUSE; NAVIGATION; POLAR NAVIGATION.

[A. B. MOODY]

**Bibliography:** N. Bowditch, *American Practical Navigator*, U.S. Navy Hydrographic Office, H.O. 9, 1958; J. C. Hill, II, T. F. Utegaard, and G. Rioridan, *Dutton's Navigation and Piloting*, 1958; U.S. Navy Hydrographic Office, *Air Navigation*, H.O. 216, 1955.

## Piltdown man

A "fossil" human being, actually the product of a skillful hoax perpetrated in the early 1900s by an unknown person. Various fragments of what were believed to be two skulls were recovered from about 1908 to 1915 by Charles Dawson, a lawyer, near Fletching, Sussex, England, assisted by Dr. A. S. Woodward and Father Teilhard de Chardin. These fragments suggested a being with a thick skull but a large brain, a high forehead and an extremely apelike jaw, of early Pleistocene date. Crude stone tools and fossils of several Pleistocene mammals were associated. Accepted by some anthropologists as an actual primitive man, but assumed by others to be the accidental mixture of human and anthropoid remains, Piltdown man caused great difficulty in interpretations of human evolution until the hoax was detected and exposed in 1953 by Prof. Sir W. LeGros Clark and Drs. J. S. Weiner and K. Oakley. They established, by chemical and other tests, that the jaw was actually that of an orangutan, modern, and not fossilized but stained to appear so; that the teeth had been filed to change their appearance; that the rest of the skull was chemically different from the jaw and probably pathological; and that the fossil animals had been introduced from various other places.

The fossil was named *Eoanthropus dawsoni* by Woodward, 1912 (also *Homo sapiens dawsoni* by Kleinschmidt). The jaw alone was named (by those originally assuming it to be an ape's) *Pan vetus* (Miller, 1915), and *Boreopithecus dawsoni* (Friederichs, 1932). The remains are at the British Museum of Natural History. [W. W. HOWELLS]

**Bibliography:** J. S. Weiner, *The Piltdown Forgery*, 1955.

## Pimento

A type of pepper (*Capsicum annum*) grown for its thick, sweet-fleshed red fruit. A member of the plant order Tubiflorales, pimento is of American origin, and gets its name from the Spanish word designating all sweet peppers. In the United States, however, the term pimento generally refers to the heart-shaped varieties grown in the South for canning and used for stuffing olives and flavoring foods. Perfection is a popular variety. Harvesting begins when the fruits are fully red, usually 2½–3 months after planting. Georgia is the only important producing state. The total annual farm value in the United States is approximately \$2,000,000. See PEPPER; TUBIFLORALES; VEGETABLE GROWING.

[H. J. CAREW]

## Piña

Piña, or pineapple fiber, is obtained from the large leaves of the pineapple plant grown in tropical countries. This natural fiber is white and especially soft and lustrous. In the Philippine Islands, it is woven into piña cloth, which is soft, durable, and resistant to moisture. Piña is also used in making coarse grass cloth and for mats, bags, and clothing. See FIBER, NATURAL; PINEAPPLE. [M. D. POTTER]

## Pinch effect

A name given to manifestations of the magnetic self-attraction of parallel electric currents having the same direction. Since 1952, the pinch effect in a gas discharge has become the subject of intensive study in laboratories throughout the world since it presents a possible way of achieving the magnetic confinement of a hot plasma (a highly ionized gas) necessary for the successful functioning of a thermonuclear or fusion reactor.

**Ampere's law.** The law of attraction which describes the interaction between parallel electric currents was discovered by André Marie Ampère in 1820, and can be stated as follows: the force of attraction in dynes per centimeter length between two thin straight wires  $r$  cm apart carrying currents of  $I_1$  and  $I_2$  amperes (amp), respectively, is  $I_1 I_2 / 100r$  (see AMPERE'S LAW). The law applies equally to the attraction between the individual components of a current in a single wire, in which case, for a cylindrical wire of radius  $r$  cm carrying a total current of  $I$  amp, it manifests itself as a compression force on the material of the wire (Fig. 1), given by  $I^2 / 200\pi r^2$  dynes/cm<sup>2</sup>. For a uniformly distributed current in the wire, the pressure reaches this value on the axis.

For the electric currents of normal experience, this force is small and passes unnoticed, but note that the pressure increases with  $I^2$ . At 100,000 amp, the pressure amounts to about 1 atm for a wire of 1-cm radius, but at 10<sup>8</sup> amp, the pressure is about 10<sup>6</sup> atm, which is considerably greater than the pressure produced by the detonation of trinitrotoluene (TNT).

**Manifestations.** The pinch effect first showed up practically in certain early types of induction

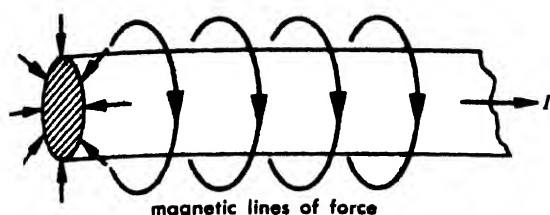


Fig. 1 Pinch pressure on a current-carrying conductor. Arrows at left show direction of pinch pressure.

electric furnaces in which large low-frequency alternating currents of the order of 100,000 amp were induced at low voltage into a horizontal ring-shaped fused-metal load (see Fig. 2). At these currents, the pinch pressure can be larger than the hydrostatic pressure exerted by the fused metal, and as the formula above shows, the pinch pressure increases as the radius of the conductor decreases. Consequently, once the process starts, the pressure at a narrow neck in the ring of fused metal is able to squeeze out the fluid metal until the neck pinches off completely, cutting off the current. This led to very uneven heating of the charge. The term pinch effect was given to this process by C. Hering in 1907. The technical difficulty was eventually overcome by making the plane of the ring vertical and submerging it deeply below the free surface of the fused metal. The force of the pinch effect has also been known to manifest itself by a crushing of tubular conductors exposed to large impulsive currents such as occur in lightning strokes or high-power short circuits.

**Thermonuclear applications.** The temperatures required for a profitable energy balance from the thermonuclear burning of deuterium are in the region of  $50 \times 10^6$ °K to several hundred  $\times 10^6$ °K. Such a temperature could be attained and maintained only if the heated plasma were carefully insulated on all sides from the walls of the apparatus by a vacuum which has the function not only of preventing the plasma from being cooled by touching the walls, but also of excluding from the plasma any foreign matter, which would inevitably be evaporated from any known material which was exposed to such temperatures. The outward pressure of the plasma must be balanced in some way without resting on material walls, and this has to be done by some combination of electric and magnetic fields. See FUSION, NUCLEAR; MAGNETOHYDRODYNAMICS; PLASMA PHYSICS; THERMONUCLEAR REACTION.

There are only a limited number of ways in which a magnetic field can be arranged around the plasma to hold it together, and one of these is that of the pinch effect. A fusion reactor using this type of confinement would ideally be a toroidal tube in which the confined plasma would float, the plasma carrying a large electric current induced in it by magnetic induction from a transformer core passing through the axis of the torus. The fundamental equation for the pinch effect in a gas, derived theoretically by W. Bennett in 1934, gives the current  $I$

required for the inward pinch pressure to balance the outward gas pressure:

$$I^2/200 = Nk(T_e + T_i)$$

where  $I$  is the total current in amperes,  $N$  is the number of electrons (also the number of ions) per centimeter length of the pinch (independent of the radius),  $k = 1.4 \times 10^{-16}$  erg/°K (Boltzmann's constant), and  $T_e$  and  $T_i$  are the temperatures in °K of the ions and electrons, respectively.

**Experimental studies.** In general, two types of apparatus have been used in studies of the pinch effect: (1) straight discharge tubes composed of quartz or porcelain with a metal electrode at each end, intended for short-duration studies, where the cooling of the plasma by the relatively cold electrodes is slight during the time of the experiment, and (2) toroidal discharge tubes, also composed of quartz or porcelain, where the pinch is endless and consequently is more effectively confined than in the first type of apparatus, and the current is induced into the discharge by magnetic coupling to a primary winding. In both cases, currents of 50,000–500,000 amp are obtained with gradients of 10–100 volts/cm along the pinch. The primary power source is a charged condenser with a capacity of 4–50  $\mu$ f, charged to 10–100 kv. The current in the discharge rises rapidly, reaches a peak in a few microseconds, and decays to zero in a damped oscillation.

**Instability.** Characteristically, as can be seen by high-speed photography, the discharge forms at the inner surface of the discharge tube wall and contracts inwardly, forming an intense line on the axis (Fig. 3); the incoming wave usually bounces slightly; the contracted discharge rapidly develops necks and kinks, and in a few microseconds, all structure is lost in an apparently turbulent glowing gas. Thus, the pinch turns out to be unstable, and the plasma confinement is soon lost by contact with the wall. The cause of the instability is easily seen qualitatively; the pinch confinement can well be described as being caused by the magnetic lines encircling the pinch behaving as slippery rubber bands which are stretched longitudinally, but which transversely are in compression (see Fig. 4). For a uniform cylindrical pinch, the magnetic pinch pressure is everywhere equal to the outward plasma pressure, but at a neck or on the inward side of a kink, the magnetic lines crowd together, creating

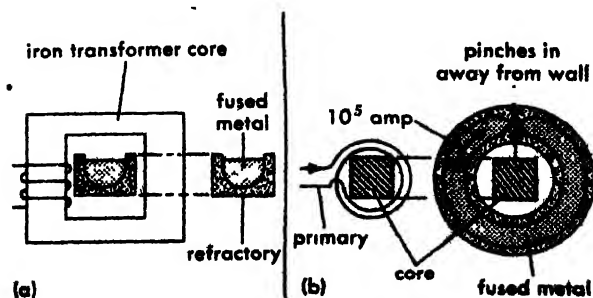


Fig. 2. Ring induction furnace. (a) Side view. (b) Plan view.



a higher pressure than the outward gas pressure. Consequently, the neck contracts down further, and the kink cuts in on the concave side and bulges out on the convex side. The effect is basically the same as that in the furnace (Fig. 2).

Neutrons are reported from linear pinch machines sometimes in great numbers (nearly  $10^9$  per pulse from the high-power Los Alamos linear pinch machine known as Columbus II). It turns out that these neutrons are emitted preferentially in the direction in which a deuteron would be accelerated by the applied electric field and are associated in some way with the instabilities that have been mentioned. They are not produced by a thermonuclear (fusion) reaction occurring throughout the pinch.

The instability has a disastrous effect on the achievement of thermonuclear temperatures by the pinch effect, and great effort has been devoted to overcoming it.

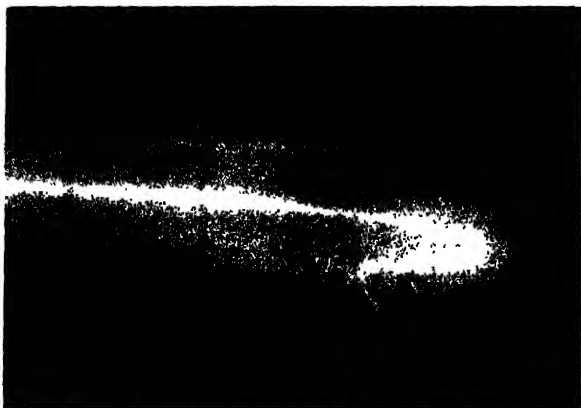


Fig. 3. Xenon pinched discharge in Perhapsatron torus.

One measure undertaken was to add an axial magnetic field by means of an external winding round the pinch tube. This might be expected to resist the sausage and kirk deformations by stiffening the discharge. Also, the walls of the tube can be made highly conducting; this step has the effect of trapping the magnetic field between the pinch and the wall, cushioning and reflecting the moving pinch back to the center. Studies involving such modification were in progress on a world-wide scale over the period 1955–1963, notably including the Perhapsatron (Los Alamos, U.S.A.), ZETA (Harwell, U.K.), and ALPHA (Leningrad, U.S.S.R.). In general, these measures proved inadequate to give stability, and work on pinches of this type has declined.

**Levitrons.** A much more powerful measure for stabilizing the pinch has been found by adding (1) a stiff current-carrying conductor down the axis inside the pinch and (2) a strong longitudinal magnetic field from an exterior winding outside it. The plasma is, in effect, sandwiched and pinched into a tubular region, between magnetic fields having directions differing by  $90^\circ$ . This so-called

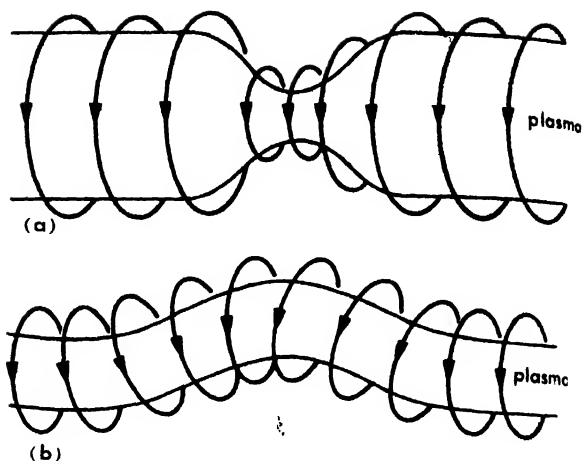


Fig. 4. Illustrations of instability. (a) Sausage type. (b) Kink type.

hard-core pinch shows greatly diminished instability, so much so that for short time scales ( $\sim 10^{-8}$  sec) and straight tubes, it is stable. In order to extend this study to longer times, the tube would have to be made impractically long to avoid cooling of the plasma by the ends or to be made endless, that is, toroidal. The achievement of a toroidal hard core involves the unusual expedient of magnetically levitating the hard-core center ring conductor (hence the name *Levitron*, applied to the device), since it cannot be reached by supports or conductors without contacting the plasma, which would cool and contaminate it (Fig. 5). Several Levitrons are built or under construction, notably, Livermore (U.S.), Aldermaston (U.K.), and Fontenay-aux-Roses (France).

The  $90^\circ$  shear in the direction of magnetic field lines in crossing the hard-core pinch plasma is the basis of the stability. Under the condition that magnetic lines cannot cross one another, plasma deformations involve much stretching of the magnetic lines and become energetically impossible. The condition that magnetic lines cannot cross holds true, however, only in a time scale depending

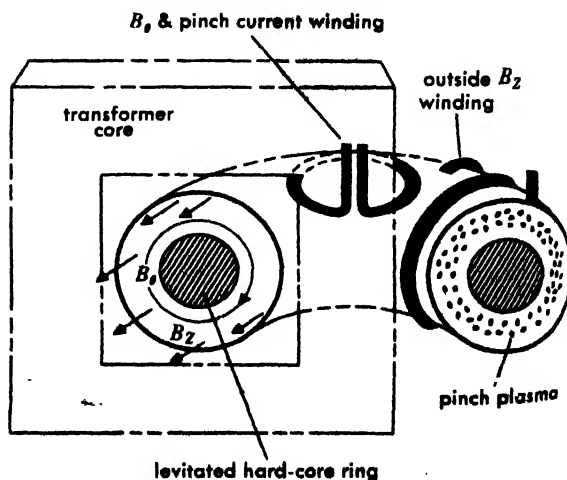


Fig. 5. Levitron or toroidal hard-core pinch.



on the electrical conductivity of the plasma, with infinite conductivity (infinite temperature) corresponding to infinitely slow crossing. Unfortunately, it has not been possible to heat the plasma in the Levitron to a high enough value to make this a good approximation for time scales larger than  $10^{-6}$  sec, and an instability is observed after such times. So far, temperatures in the thermonuclear range have not been reached in the hard-core pinch, and its future in the controlled-fusion effort is not yet clear. The hard core is associated with S. A. Colgate, H. P. Furth, and P. Rebut.

A loophole still remains for the utilization of the simple pinch in controlled fusion—that of resorting to heroic extremes of high power, short time (below that required for instabilities to develop), and high plasma density; and it is still being studied with this possibility in mind. [J. L. TUCK]

**Bibliography:** C. L. Longmire, J. L. Tuck, and W. B. Thompson (eds.), *Plasma physics and thermonuclear research*, *Prog. Nucl. Energy, Ser. XI*, 1963; Theoretical and experimental aspects of controlled nuclear fusion, in *Proc. Second Intern. Conf. Peaceful Uses At. Energy*, vols. 31–32, 1958.

## Pine

The genus, *Pinus*, of the pine family, characterized by evergreen leaves, usually in tight clusters (fascicles) of 2–5, rarely single. There are about 80 known species distributed throughout the Northern Hemisphere. Botanically, the leaves are of two kinds: (1) a scalelike form, the primary leaf, which subtends a much shortened and eventually deciduous shoot bearing (2) the secondary leaves or needles. The wood of pines is easily recognized by the numerous resin ducts and by the characteristic resinous odor (see SECRETORY STRUCTURES, PLANT). The pines may be divided into two classes according to the number of leaves in a cluster.

**Soft or white pines.** Except for the nut pines, the white pines have 5 needles in a cluster. Eastern white pine, *P. strobus*, ranging from 90 to 150 ft in height, is found in the northeastern United States west to the Lake states, adjacent Canada, and the Appalachian Mountain region. Originally it was the most important timber tree of the eastern United States and Canada, with a stand of approximately 750,000,000,000 board ft. About 800,000,000 board ft are cut annually, and the total quantity of saw timber in the United States at present is estimated at about 15,000,000,000 board ft. In the amount of annual lumber cut, eastern white pine is now exceeded by the southern or hard pines, Douglas-fir, oak, ponderosa pine, and hemlock (see DOUGLAS-FIR; HEMLOCK; OAK). The wood of eastern white pine is valuable because it can be easily worked, is light and soft, does not split when nailed, polishes well, and does not warp or swell appreciably. Almost everything from shipmasts to matches, including doors, framing, finish, boxes, and crates, has been made from this wood. However, it is now restricted to more particular uses because of its increasing scarcity and value.

Of the western white pines, the sugar pine,

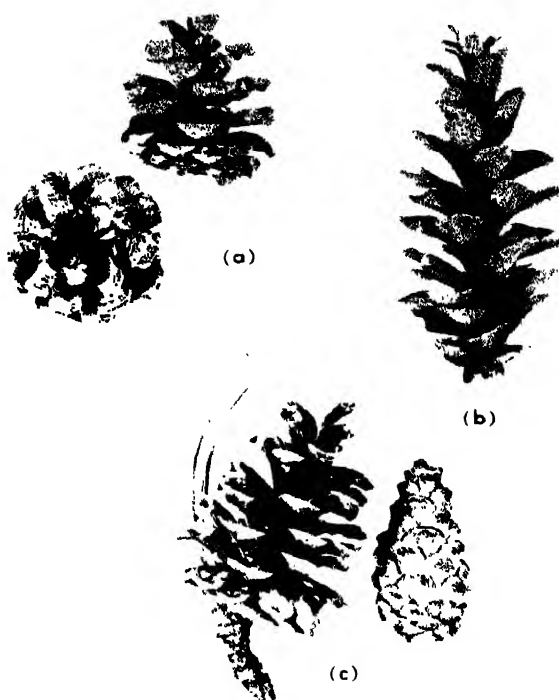


Fig. 1. Pine cones. (a) *Pinus resinosa*. (b) *P. strobus*. (c) *P. nigra*. (Brooklyn Botanic Garden, Brooklyn, N.Y.)

*P. lambertiana*, is the most important. This is a magnificent Pacific Coast tree attaining a height of about 250 ft and having large cones, 10–20 in. long. It is lumbered only in Oregon and California where it ranks in volume and value with redwood (see REDWOOD). The annual cut is 300,000,000–400,000,000 board ft. The present stand is estimated at about 20,000,000,000 board ft, with approximately four-fifths in California. Other white pines in the West are western white pine, *P. monticola*, a mountain species found almost entirely in Idaho, Montana, and Oregon; limber pine, *P. flexilis*, one of the smaller white pines of the Rockies; and white-bark pine, *P. albicaulis*, also a mountain species with a more northern range.

The nut pines or piñons are a subgroup of the Southwest with fewer needles, sometimes only one.

**Hard or pitch pines.** Red pine, *P. resinosa*, also known as Norway pine, reaches a height of 80 ft or more, and is native in the northeastern United States from Maine to Minnesota and adjacent Canada and south along the mountains to West Virginia. The needles are in pairs, and the bark has a red-brown color, hence the name. The wood is fairly soft, but a little harder than that of eastern white pine. The more dense red pine is also stronger and is important commercially for general construction, sash, door, and window frame manufacturing, flooring, boxes, crates, and shipmasts, but it is not durable in contact with the soil. Frequently it is sold in mixture with eastern white pine. The stand in the United States has been estimated at about 3,000,000,000 board ft with about two-thirds of the stand in the Northeast and the remainder in the Lake states. The average annual cut is about 150,-

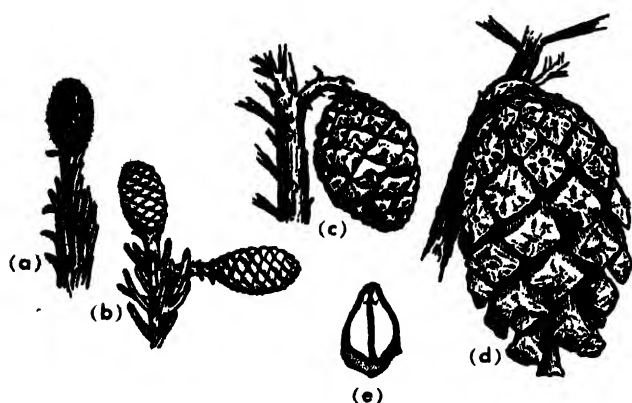


Fig. 2. Scotch pine, *Pinus sylvestris*. (a) Female (ovulate) cone unfertilized. (b) Fertilized cones. (c) Stage in development of cone. (d) Cone opening to discharge seeds. (e) Seed scale showing two winged seeds. (A. H. Graves, *Illustrated Guide to Trees and Shrubs*, Harper, 1956)

000,000 board ft. Technologically, the red pine forms a sort of bridge between the soft white pines and the hard pines.

The hard yellow pines have two or three needles in a cluster. The longleaf pine, *P. palustris*; loblolly, *P. taeda*; shortleaf, *P. echinata*; and the slash pine, *P. elliottii*, are the principal trees of the southeastern United States that yield the so-called southern yellow pine lumber, a hard, resinous wood in which the dark bands of summer wood conspicuously alternate with the lighter colored spring wood. See WOOD (ANATOMY AND IDENTIFICATION). The longleaf pine is the most important of these, being a leading world producer of naval stores. The hard, strong wood is used in construction, flooring, railway car construction, and shipbuilding. The longleaf and slash pines are the most important trees for the production of turpentine, a distillate obtained from the resin when the trees are tapped. See GUM; NAVAL STORES; ROSIN.

In the West the hard pines are represented chiefly by the ponderosa pine, *P. ponderosa*, which attains a height of 150–225 ft, and is found in the Rocky Mountain and Pacific Coast regions, including adjacent Canada. The total stand and lumber cut from this pine (including Jeffrey pine, *P. jeffreyi*, which is commonly sold along with it) are exceeded only by Douglas-fir and the southern yellow pines. The total stand of saw timber is about 185,000,000,000 board ft and the annual cut between 3,000,000,000 and 4,000,000,000 board ft, nearly half of which comes from Oregon and about one-quarter from California.

The Austrian pine, *P. nigra*, which has two needles and closely resembles the red pine, has darker bark and is much planted in North America, as is the Scotch pine, *P. sylvestris*, with two shorter, bluish needles (Fig. 2). Other exotics cultivated are mountain pine, *P. mugo*; Japanese black pine, *P. thunbergi*; and Swiss stone pine, *P. cembra*, the last a 5-needled species. See FOREST AND FORESTRY; TREE; TURPENTINE.

[A. H. GRAVES]

## Pine oil

A material fractionated from oils recovered from long leaf pine wood (see WOOD CHEMICALS). Pine wood is either destructively distilled or solvent extracted to yield turpentine, rosin, charcoal, and other useful commercial products. The recovered oils are further refined, and pine oil results from a definite cut in the fractionation.

Good grades of pine oil consist largely of terpineol. A typical pine oil contains

	%
$\alpha$ -Terpineol	65–70
Dihydro- $\alpha$ -terpineol and other tertiary alcohols	10
Borneol and fenchyl alcohols	10–15
Estragole	5
Ketones	5–10

The textile industry uses pine oil as a penetrant, dispersing agent, wetting agent, and inhibitor of bacterial growth in practically all wet processing of cotton, silk, rayon, and woolen goods. In the paper industry it is used as a wetting and leveling agent in coating and as a preservative of casein. It finds extensive use for the preparation of high-grade disinfectants, liniments, odorant blocks, dog soaps, cattle sprays, and other insecticidal preparations. Manufacturers of essential oils use pine oil as a source of prime terpineol and also of some other constituents. About 28,000,000 gal of pine oil is produced annually in the United States.

A number of essential oils are derived from the needles and cones of other species of pines, most of them produced in Switzerland, Austria, Germany, and Russia. See ESSENTIAL OILS. [E. L. SAU]

## Pineal body

In its most complete form, the pineal may be regarded as consisting of a series of evaginations or outgrowths from the third ventricle in the roof of the diencephalon of the brain. The pineal apparatus constitutes one of the oldest parts of the brain and is widespread throughout the vertebrate animal kingdom. The two principal components of this apparatus are usually known as the parietal eye or pineal eye and the pineal organ or the epiphysis. See BRAIN. [H. J. CLAUSEN]

**Comparative anatomy.** The pineal organ is an unpaired, elongated, club-shaped, knoblike or occasionally threadlike organ attached by a stalk to the roof of the forebrain. In lower vertebrates it often projects upward through the skull to lie under the skin; in higher ones it is hidden underneath the enlarged cerebral hemispheres. No evidence of the organ has been found in the simplest cyclostomes (myxinioids), and it is said to be lacking in crocodiles, armadillos, and edentates. Embryonic vestiges occur in some sirenians. It is small in birds and some mammals, microscopic in the porpoise, and relatively large in cocks, echidnas, marsupials, rodents, ungulates, and man. Nervous elements predominate in lower vertebrates, while glandular cells predominate in higher ones.

**Lower vertebrates.** The pineal body of lower fishes may shed light on its phylogenetic history. Two outgrowths project from the dorsal forebrain stalk in cyclostomes. One, derived from the right side of the brain, is the pineal body; the other, derived from the left, is the parapineal. The pineal ends in a hollow, knoblike termination directly beneath an area of skin devoid of pigment and lying dorsally between the paired eyes. The upper wall of the organ consists of several layers of cells, forming a lens. The lower wall contains sensory cells and ganglion cells with processes passing down the stalk to nerve centers in the right side of the brain. The parapineal is less specialized but essentially similar in structure and is connected to the left side of the brain. Both organs are sensitive to light. In some ganoids and lower teleosts all tissues overlying the pineal are translucent, and light striking the pineal initiates reflex responses of locomotor muscles and pigment cells.

**Higher vertebrates.** In modern fishes and tetrapods the pineal is chiefly glandular and neuroglial and is confined within the brain case, but non-functional sensory and ganglion cells remain, the latter even in mammals. In birds, the pineal shows little nervous or glandular structure.

In certain reptiles (*Sphenodon*, some lizards), a parapineal (parietal) eye, less specialized than that of cyclostomes, is exhibited in addition to a pineal body. A parapineal anlage occurs in amphibian and bird embryos, and an adult parapineal remains in certain extant, but mostly ancient, fishes. See SENSE ORGAN.

Evidence, not all of which is summarized here, suggests that the pineal might represent one of a pair of phylogenetically older sense organs, the other being the now almost obliterated parapineal. Not all authorities agree; however, if this is true, the pineal has slowly evolved from a photoreceptor to an organ which subsequently produced, and may still produce, an endocrine substance.

[G. C. KENT, JR.]

**Physiology.** The function of the pineal body has not been definitely ascertained. The pineal organ or epiphysis is that component of the pineal apparatus which appears to be glandular in most vertebrates, including man. It has been shown to be secretory in some of the lower vertebrates.

Claims have been made, particularly by English scientists, that in mammals the pineal delays the onset of sexual maturity and thus acts as an antagonist to the gonadotropic hormones of the hypophysis. See ENDOCRINE GLAND; HORMONE.

The pineal apparatus is in itself highly vascular, so that its secretions could easily enter the blood stream.

**Parietal eye.** In many species, particularly among many of the cold-blooded animals, the parietal eye component is present and in a position to be influenced by solar radiation. In some species it forms a vesicle situated beneath a layer of translucent skin; in others, it lies within or beneath a light-transmitting region of a foramen in the skull.

The parietal eye component may or may not have a direct nerve connection with the brain. When a nerve connection is evident, it connects with a region of the brain known as the habenular apparatus, which in turn is known to be connected to such structures as the hypothalamus and midbrain as well as to the internal capsule in some species.

**Function.** Experiments, chiefly on fishes and amphibians, indicate that there is an influence on melanic pigment dispersal and response to light. Studies on reptiles suggest a role in metabolism and reproduction, and experiments on birds and many mammals also indicate some effects on the reproductive system. Clinical studies in man also suggest that some physiological relationship between the pineal and the reproductive system exists.

The intimate association with important neural pathways, the abundant vascular supply, and the cytological evidence, as well as the recent studies with radioactive isotopes, suggest a high level of physiologic activity for the pineal organ or epiphysis. Studies with regard to light reception in the parietal eye component in cold-blooded vertebrates suggest some sensory role with respect to solar radiation.

[H. J. CLAUSEN]

**Bibliography:** R. J. Gladstone and C. P. G. Wakeley, *The Pineal Organ*, 1940; J. I. Kitay and M. D. Altschule, *Physiology of the Pineal Gland*, 1954.

## Pineapple

A low-growing perennial plant, indigenous to America. The cultivated varieties belong to the species *Ananas sativus*, of the plant order Farnales.

The edible portion of the pineapple develops



Pineapple, *Ananas sativus*, showing "fruit" and leaves (USDA)

from a mass of ovaries on a fleshy flower stock having persistent bracts. On the cultivated types, the flowers are usually abortive. The leaves are long and swordlike, usually rough edged, and grow to a height of 2-4 ft. Commercial plantings bear fruit at the age of 12-20 months, and may continue to be productive for as much as 8-10 years. Propagation is by suckers or offsets which may be rooted in sand, but are usually set directly in the field where they are to produce. The leafy crowns may also be used as cuttings, but because they are harvested with the fruit, other methods of propagation are more satisfactory. See FARINALES.

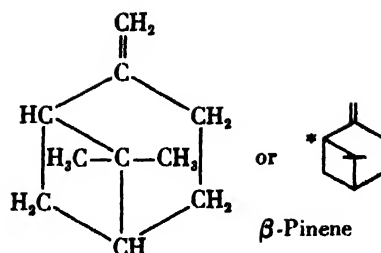
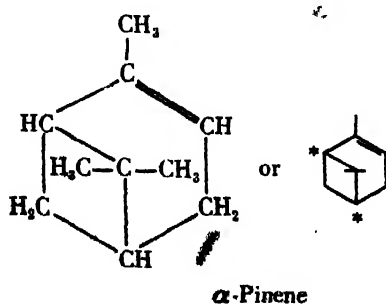
The pineapple, a warm-climate plant, is injured by temperatures below 32°F. It does best in a dry atmosphere and relatively poor soil, but responds well to fertilizers. The major producing area is the Hawaiian Islands where special methods of culture and harvesting have been developed. A paper mulch is used in the production of much of the Hawaiian pineapple crop. The paper is laid in long strips by a special machine which pulls soil over the edges of the strips to hold them in place. The mulch aids in weed control and decreases moisture evaporation from the soil. Because of the tendency to iron-chlorosis of plants growing on these soils, spraying with an iron salt is frequently a part of the fertilization program. Careful control of the nitrogen supply in relation to the hours of sunshine, together with the use of hormone sprays under certain conditions, has made possible considerable control over the time of ripening, an important consideration in obtaining maximum year-round use of processing facilities. Pineapples are also grown in the West Indies and to a limited extent in southern Florida. See FRUIT GROWING (SMALL).

Pineapples are consumed fresh in considerable quantity, but because of distance from markets and the problems of transporting fresh fruit, most of the crop is canned as sliced pineapple or as juice. In Hawaii, the average annual value of the processed crop was about \$91,000,000 in recent years.

[J. H. CLARKE]

## Pinene

One of the most important of the terpene hydrocarbons.  $\alpha$ -Pinene is widely distributed in nature and is the chief constituent of oil of turpentine. The beta variety often accompanies the alpha, and is usually present in small quantities. Both exist as the dextro, levo, and racemic forms of optical isomers. Although theoretically  $\alpha$ -pinene should be



capable (starred carbons) of having two dextro- and two levorotatory varieties, no evidence of such isomerides has been demonstrated. See TURPENE.

The pinenes are usually isolated by fractional distillation from essential oils, although the  $\alpha$ - and  $\beta$ -pinenes are difficult to separate from each other.

Both pinenes are colorless oils which resinify on exposure to air. They are soluble in most organic solvents. As bicyclic terpenes, they readily undergo molecular rearrangements forming products having a changed ring structure.

### Physical properties

	$\alpha$ -Pinene d-, l-, and dl-	$\beta$ -Pinene d-, l-
Boiling point, 760 mm.	156°C	164-166°C
Density (20°C)	0.858-0.860	0.87
Refractive index, 20°/D	1.466	1.46-1.48
Specific rotation	d = +51° l = -51°	l = -22°

$\alpha$ -Pinene is the starting point in the commercial production of borneol, camphor, terpineol, and terpin hydrate. See TURPENTINE. [E. L. SAU]

## Pink eye

An acute, contagious conjunctivitis caused by the bacterium *Haemophilus aegyptius* (Koch-Weeksbacillus). This organism closely resembles *H. influenzae* but is unencapsulated and serologically different. Pink eye may occur in epidemics, particularly among children in warm climates. Diagnosis is based upon the characteristic diffuse inflammation of the conjunctivae and sclera and upon the isolation of *H. aegyptius* on a blood-containing culture medium such as chocolate agar. It is treated with neomycin locally or broad-spectrum antibiotics orally. Recovery produces little immunity. See HEMOPHILIC BACTERIA. [W. F. VERWEY]

## Pinta

A disease similar to syphilis caused by a spirochete which occurs principally in certain rural tropical areas of Central and South America; cases have been described from other tropical regions. The early lesions and the general evolution of pinta are similar to, but generally milder than, those of syphilis and yaws. Generalized lesions develop a reddish-purplish hue which is followed by atrophy and spotty achromia of the skin. Treponemes, morphologically similar to *Treponema pallidum*, are present in small numbers in the skin lesions. The infection has not been established in laboratory animals but has been transmitted to human volunteers. Ser-

ous late complications have not been described. The serological pattern and response to treatment are similar to syphilis. Transmission is presumably by direct contact. See SPIROCHETE; SYPHILIS; YAWS. [T. B. TURNER]

## Pinworm

A species, *Enterobius vermicularis*, of the class Nematoda, phylum Aschelminthes, also called seatworms. This worm is one of the most common and annoying parasites of man. Some estimates of the degree of infestation in the United States have run as high as 35% of the total population. It is especially prevalent in children. In some communities the incidence of infection approaches 100%.

The anatomy of the pinworm is basically similar to that of *Ascaris*. It is a small, white worm, the males being only 2-5 mm long, the females 9-12 mm long. The posterior third of the body of the male is curved into a tight spiral, in contrast to the straight, long, pointed tail of the female.

Heavy infestations by this worm are common. It lives in the upper portion of the large intestine. The gravid females descend into the rectum and pass out with the feces, or they may crawl out of the rectum at night to deposit their eggs over the perianal area.

The life cycle is complete within an individual. The infection builds up by the transmission of eggs by the fingers, fingernails, and bed clothing to the mouth. Each generation lasts only 3-4 weeks and the infection will soon die out if reinfestation by the ingestion of eggs is prevented.

Slight infections may pass unnoticed, but heavier infections will cause intense pruritis and intestinal disturbances. Anemia, bed wetting, and unusual restlessness in the sleep along with anal itching are usually indications of the presence of pinworms, a condition called enterobiasis. A number of chemicals kill pinworms, but rigid precautions against reinfestation must be taken.

Nose picking by children sometimes results in infection of the nasal mucosa. The worms, frequently found in the vermiform appendix, may be the cause of a substantial percentage of the cases of appendicitis. Pinworms occur throughout North America, but they are more prevalent in warm areas. See ASCARIDIDA; ASCARIS. [J. D. BLACK]

## Pinworm infection

An infection, also known as enterobiasis, of the human intestinal tract by the pinworm, *Enterobius vermicularis*. The female nematode is  $\frac{1}{4}$  in. long, the male much smaller. Both inhabit the large intestine but the vagina may also be invaded. The female produces several thousand eggs which accumulate in its uteri. Pregnant worms, 2 to 4 weeks old, are expelled in the feces, or crawl out of the anus to lay the eggs and die. Their movements cause itching, and nervous disorders may ensue.

Scratching contaminates clothing and hands; the patient may thus become autoinfected and, in addition, scatter eggs in the house. Thus the incidence is high in children's institutions in temper-

ate climates. Retroinfection may occur by the larva's breaking out of the egg and entering via the anus to develop.

The disease is diagnosed by observing the worm upon extrusion from the anus and by microscopical detection of eggs in anal swabs. It is treated by oral administration of piperazine or dithiazanine for 1 week. For taxonomy, see PARASITOLOGY, MEDICAL; STRONGLYOIDEA. [J. F. MALDONADO]

## Pipe flow

Conveyance of fluids in closed conduits. Pipes have been used for thousands of years, but the detailed understanding of velocity distributions and of energy losses has come about during the twentieth

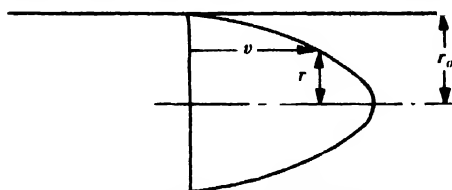


Fig. 1. Laminar velocity distribution.

century. Satisfactory equations for predicting the losses in flow of water have been in use since about the middle of the nineteenth century.

**Laminar flow.** The fluid moves parallel to the pipe axis in a straight, round pipe when the Reynolds number  $VD/\nu$  is less than 2000, in which  $V$  is the average velocity,  $D$  the pipe diameter, and  $\nu$  the kinematic viscosity (see REYNOLDS NUMBER). The velocity distribution in laminar flow is parabolic (Fig. 1) and the equation for velocity  $v$  is

$$v = \frac{\Delta p}{4\mu L} (r_0^2 - r^2)$$

in which  $\Delta p$  is the pressure drop in length  $L$ ,  $\mu$  is the absolute viscosity, and  $r_0$  and  $r$  are as shown in Fig. 1. Discharge  $Q$  in laminar flow is

$$Q = \frac{\Delta p \pi D^4}{128\mu L}$$

and is independent of the roughness of the interior pipe wall surface.

In laminar flow the fluid particles move in straight lines parallel to the pipe axis, in telescoping layers with each inner layer moving more rapidly than its adjacent outer layer. Energy losses vary as the first power of the velocity, so that by doubling the discharge (or average velocity) the pressure drop is doubled.

**Turbulent flow.** In a pipe when the Reynolds number is greater than 2000-4000 normally the fluid particles no longer move parallel to the pipe axis. The exact transition velocity depends upon the nature of the piping system. For high Reynolds numbers, the orderly motion of laminar flow becomes unstable, with fluid particles moving in random paths with large transverse velocity components. Energy is dissipated in the turbulent motion, with the loss varying as the velocity to the 1.7-2.0 power. The Darcy-Weisbach equation for head loss  $h_f$  due to turbulent flow in a pipe is

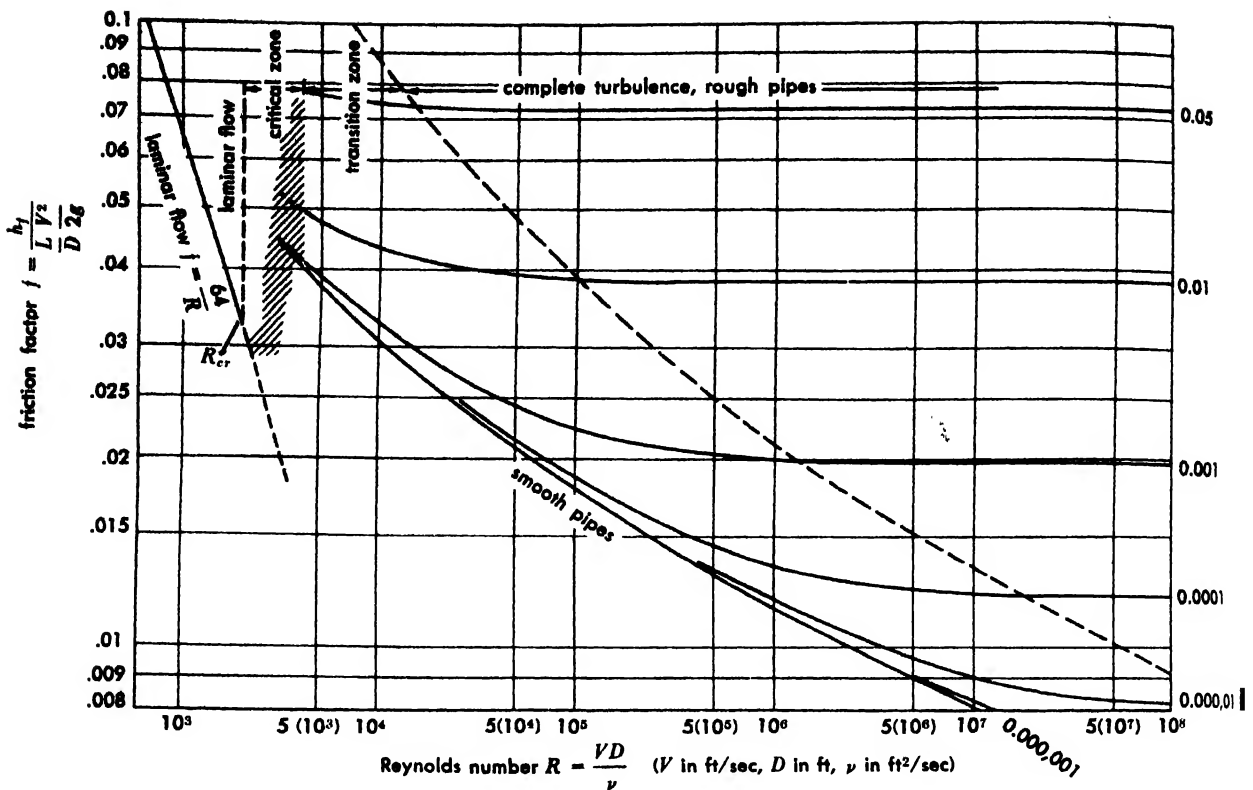


Fig. 2. Moody diagram.

$$h_f = f \frac{L V^2}{D 2g}$$

in which  $V$  is the average velocity,  $L$  is the length,  $D$  the diameter, and  $f$  a dimensionless factor dependent upon the wall roughness, the fluid properties, and upon the velocity and pipe diameter  $f = f(V, D, \rho, \mu, \epsilon)$  with  $\rho$  the fluid density and  $\epsilon$  a measure of the absolute roughness of the pipe wall, having the dimensions of a length.

For turbulent flow in smooth pipe,  $\epsilon = 0$  and the expression for  $f$  becomes  $f = f(VD\rho/\mu)$  in which  $VD\rho/\mu$  is Reynolds number  $R$ . The form of the functional relation between  $f$  and  $R$  must be determined by experiment, and is shown as the lowest curved line on the Moody diagram (Fig. 2).

Laminar flow may also be shown on the Moody diagram, because its equation may be written

$$h_f = \frac{64 L V^2}{R D 2g}$$

For rough pipes  $f = f(R, \epsilon/D)$  in which  $\epsilon/D$  is known as the relative roughness. An empirical equation has been worked out by C. F. Colebrook which is the basis for the Moody diagram. It gives good results for new commercial pipes, with values of  $\epsilon$  as shown in the left-hand lower corner of the chart.

To find the head loss for flow of a given amount of liquid per unit time through a pipe of known size, length, and type of manufacture, the Reynolds number and the relative roughness are computed and then used in the Moody diagram to determine  $f$ . With  $f$  known, all quantities in the Darcy-Weis-

bach equation are known except  $h_f$ , so it can be determined.

When the amount of head loss is known but the discharge (volume per unit time flowing) is desired, a trial solution is required. An  $f$  is assumed from the Moody diagram for the known  $\epsilon/D$ , and by its use a trial value of  $V$  is found from the Darcy-Weisbach equation. With this  $V$ , a trial Reynolds number is computed which permits a better value of  $f$  to be found from the Moody diagram.

With flow of a gas, the same methods may be used as with a liquid if the change of density is small (less than 10%). For large density changes the equation of state relating density and pressure intensity is required, as well as special methods for obtaining head loss or weight per unit time flowing.

Head losses due to changes in direction and in size of pipe, and those due to valving are grouped as minor losses, and tend to vary as the square of velocity. They may be expressed as an equivalent length of pipe  $L_e$ , which is added to the actual length of pipe in using the Darcy-Weisbach equation.

With old pipe, wall roughness  $\epsilon'$  tends to increase linearly with time so that  $\epsilon' = \epsilon + at$  in

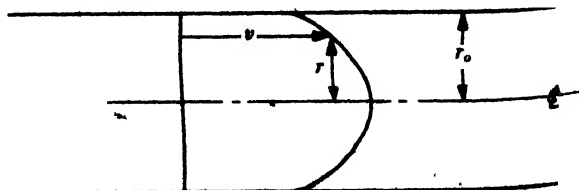


Fig. 3. Turbulent velocity distribution.

which  $\alpha$  is a constant determined by test on the particular pipe line and fluid, and  $t$  is time.

Velocity distribution in turbulent flow in a pipe is more uniform than for laminar flow, due to the large transfer of momentum radially across the flow (Fig. 3). A simple equation that gives reasonably good results is Prandtl's one-seventh power law

$$v = v_{\max} \left( \frac{y}{r_0} \right)^{1/7}$$

in which  $y$  is the distance from the pipe wall, and  $r_0$  is the pipe radius. [V.L.S.]

**Bibliography:** C. F. Colebrook, Turbulent flow in pipes, with particular reference to the transition region between the smooth and rough pipe laws, *J. Inst. Civil Engrs. (London)*, 12:133-156, 1939, V. L. Streeter, *Fluid Mechanics*, 2d ed., 1958.

## Pipeline

Major uses of pipelines are for the transportation of petroleum, water (including sewage), chemicals, foodstuffs, pulverized coal, and gases such as natural gas, steam, and compressed air. Pipelines must be leakproof and must permit the application of whatever pressure is required to force conveyed substances through the lines. Pipe is made of a variety of materials and in diameters from a fraction of an inch up to 30 ft. Principal materials are steel, wrought and cast iron, concrete, clay products, aluminum, copper, brass, cement and asbestos (called cement-asbestos), plastics, and wood.

Pipe is described as pressure and nonpressure pipe. In many pressure lines, such as long oil and gas lines, pumps force substances through the pipelines at required velocities. Pressure may be developed also by gravity head, as for example in city water mains fed from elevated tanks or reservoirs. Nonpressure pipe is used for gravity flow where the gradient is nominal and without major irregularities, as in sewer lines, culverts, and certain types of irrigation distribution systems.

Design of pipelines considers such factors as required capacity, internal and external pressures, water- or airtightness, expansion characteristics of the pipe material, chemical activity of the liquid or gas being conveyed, and corrosion.

Most pipe is jointed, although some concrete pipe is monolithically cast in place. The length of the individual sections of pipe and the method of joining them depend upon the pipe material, diameter, weight, and requirements of use. Steel pipe sections are usually joined by welding, couplings, or riveting. Cast-iron pipe may be joined by couplings or, in the case of bell-and-spigot pipe, by filling the space between the bell and the spigot with calked or melted metal such as lead. Flexible-type joints with rubber gaskets are also used for joining cast iron pipe. The rubber gasket is contained in grooves and is ordinarily the sole element making the joint watertight.

Cement-mortar-filled or lead-filled rigid-type bell-and-spigot joints are usually used for joining concrete or vitrified clay sewer pipe. Tongue-and-

groove rigid-type mortar-filled joints are often used for concrete pipe in low-pressure installations. The flexible-type joints are most frequently used for asbestos-cement pipe and concrete pipe under higher pressures. See PIPE FLOW. [L.N.M.]

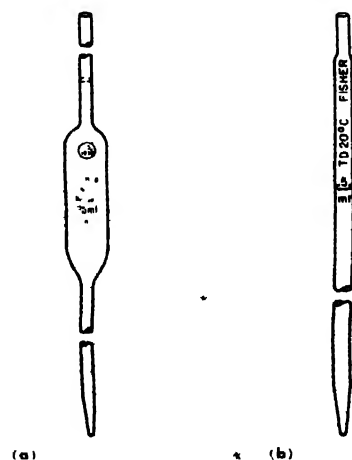
## Piperales

A primitive order of the plant subclass Dicotyledoneae including 3 families: the lizard's-tail family (Saururaceae), the pepper family (Piperaceae), and the Chloranthaceae. These plants have mainly naked flowers or, in the last family, a small bract-like perianth. Like the Saururaceae, the Chloranthaceae is a very small family of no economic value.

The pepper family includes 12 genera and 1400 species indigenous to tropical regions in both the Old and the New Worlds. Several species have considerable economic importance. The fruits of *Piper nigrum* of Malaya are dried and ground to produce commercial black pepper, or the seed only is ground to make white pepper. The dried berries of *Piper cubeba* of the East Indies are the cubebs used in medicine. The East Indian natives chew betel nuts combined with the fresh leaves of *Piper betle*, the betel pepper. See BETEL NUT; CUBEBS; PEPPER; see also DICOTYLEDONEAE; EMBRYOPHYTA; PLANT KINGDOM. [P.D.S.]

## Pipet

Pipets are usually made of glass and are used almost exclusively to deliver accurately known volumes of liquids or solutions. There are two general categories of pipets: volumetric or transfer pipets and the graduated measuring type (see illustration). Volumetric pipets are used in the following way: the liquid is sucked up above the mark on the stem above the bulb, and the upper end is quickly closed with the index finger; any adhering liquid is wiped from the outside of the lower stem; the level of liquid in the stem is allowed to fall slowly, by regulating the pressure on the finger, until the bottom of the meniscus is tangent to the mark; liquid clinging to the tip is removed; and the pipet is al-



Pipets. (a) Volumetric pipet. (b) Graduated pipet.



lowed to empty freely into the receiving vessel. After 15 sec or the time specified on the pipet for drainage, the tip is touched and rotated against the inside of the receiver. The liquid remaining in the tip thereafter is not removed. Volumetric pipets, when handled in the described manner, will deliver reproducibly a definite amount of liquid or solution.

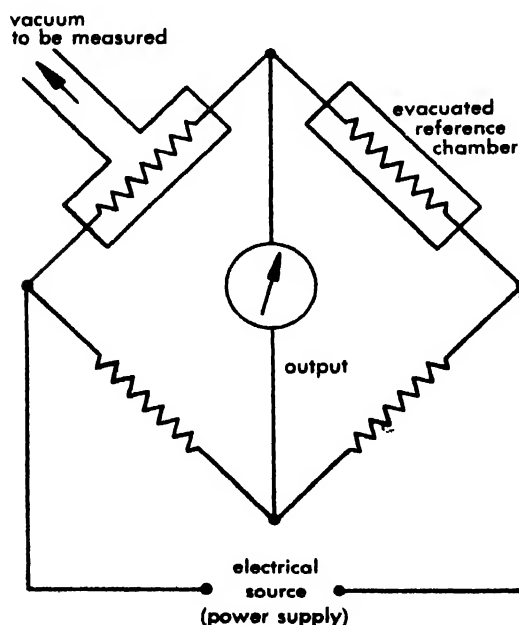
A graduated measuring pipet is used in the same general way except that the volume of liquid delivered can be varied by allowing the liquid to drain from one calibration mark to another. These pipets are usually not as accurate in delivering volumes of liquid as are the volumetric type.

For solutions that attack glass, pipets made of various plastics are used. See TITRATION; VOLUMETRIC ANALYSIS. [C.E.B.]

### Pirani gage

A type of instrument used to measure vacuum in the range of 1 micron to a few hundred microns (a micron is the pressure required to support a column of mercury 0.00004 in. high). See VACUUM MEASUREMENT.

In the Pirani gage, a fine wire filament, one of the four electrical resistances forming a Wheatstone bridge circuit, is exposed to the vacuum to be measured. Electric current heats the wire; the surrounding gas (vacuum) carries heat away from the wire. At constant pressure (vacuum) the wire quickly reaches an equilibrium temperature. If the pressure rises, the gas carries away more heat, and the temperature of the wire decreases. Since the resistance of the filament is a function of temperature, the electrical balance of the Wheatstone bridge is changed. The output meter is usually a microammeter calibrated in units of microns of pressure.



Pirani gage.

The calibration depends upon the thermal conductivity of the gas, and therefore the calibration is different for different gases. Accuracy is commonly of the order of  $\pm 5\%$  of scale.

The Pirani gage is a relatively simple and rugged instrument, widely used both in industrial plants and in laboratories. [B.D.H.; H.G.P.]

### Pisces (constellation)

The Fishes, in astronomy, is a zodiacal constellation appearing in the autumn evening sky. Pisces is the twelfth and last sign of the Zodiac. It is inconspicuous, having no star brighter than the fourth magnitude. But it is an important constellation because the vernal equinox, which marks the beginning of the astronomical year, is now located in it. Its most distinctive feature is a V-shaped figure, with the fishes' tails toward the point of the V tied together by a ribbon. The northern fish is poorly defined, but the western one is marked with a group of stars forming an irregular pentagon, known as the Cirlet in Pisces. See CONSTELLATION. [C.S.Y.]

### Pisces (zoology)

A term that embraces all fishes and fishlike vertebrates. In early zoological classifications fishes, like mammals, birds, reptiles, and amphibians, were ranked as a class of the vertebrates. As knowledge of fishes increased, it became apparent that, despite their common possession of gills and fins and their dependence on an aquatic environment, not all fishes were closely related. At least five groups of fishes with modern descendants were already established before the tetrapods appeared. Not only are these groups older, but some are decidedly more divergent structurally than are the four classes of tetrapods. For these reasons several classes of fishes are now recognized. The number of classes varies; one reputable but extreme classification scheme recognizes eleven classes of fishes.

The primary cleavage in vertebrate classification is that separating the jawless fishes, or Agnatha, from those vertebrates with jaws, the Gnathostomata. After recognition of this split, the name Pisces was commonly restricted to the jawed fishes. When these in turn were divided into two or more classes, Pisces was further restricted by some authorities to the bony fishes. Another scheme involves assignment of class names to each of the major constituent groups of jawed fishes, and use of Pisces as a superclass name. In view of the confusion, it seems best to revert to early practice and to employ Pisces as a group name of convenience to embrace all classes of fishlike vertebrates, from jawless fishes to bony fishes. In this sense it has no actual taxonomic status because it cuts across natural classification, dividing the gnathostomes and grouping part of them with the agnaths. See AGNATHA; GNATHOSTOMATA.

The Pisces includes four well-defined groups that in the light of present information merit recogni-

tion as classes: the Agnatha or jawless fishes, the most primitive; the Placodermi or armored fishes, known only as Paleozoic fossils; the Chondrichthyes or cartilaginous fishes; and the Osteichthyes or bony fishes. Future research may demonstrate the need for further division, but this is most likely to involve Paleozoic groups. See CHONDRICHTHYES; OSTEICHTHYES; PLACODERMI.

**Number of recent species.** Present fish classification is not sufficiently precise to permit an accurate tabulation of the number of living species. New kinds are constantly being discovered, others are being synonymized as the result of new research, and the literature is scattered. Nevertheless, estimates by competent ichthyologists are so diverse, ranging from 18,000 to 40,000, that an effort is here made to arrive at a reasonably acceptable approximation. The result indicates that most previous estimates are far too high. Counts for Recent groups and species include Agnatha, 2 families, about 11 genera, and approximately 45 species; Chondrichthyes, 31 families, some 132 genera, and roughly 575 species; Osteichthyes, 31 orders, about 333 families, 3,100 genera (based in part on estimates and perhaps in error by as much as 10%), and about 16,700 species. This latter figure broken down gives about 3,700 species in 29 orders, 5,000 species in the Cypriniformes, and 8,000 species in the Perciformes. The number of perciforms is an estimate based chiefly on a reasonably accurate tally of 1,200 genera and is probably too high. The true total number of Recent fish species is probably between 15,000 and 17,000.

**Ecology.** Fishes live in almost all permanent waters to which they have been able to gain access. In general they have evolved a body conformation and specialized features that adapt them harmoniously to the world about them. Inhabitants of mountain torrents may have peculiar attachment organs; those living in Antarctic waters at a temperature below freezing have made needed physiological adjustments; fishes of the deep sea commonly carry their own light source, and the female angler fish is assured a mate by the parasitism of the male on her body. In the East Indies some fishes skip with ease over mud flats, and others ascend trees. Some fishes mature at extremely small size. A Philippine goby, *Mistichthys*, reaches a length of only  $\frac{1}{2}$  in. and is commercially important as a food fish although it takes 70,000 fish to make 1 kg, and a Samoan fish, *Schindleria*, attains a weight of only 6 mg. At the other extreme, the whale shark is reputed to reach a length of 60 ft, and a 38-ft individual weighed more than 13 tons.

**Adaptive radiation.** Because their body is supported by water, fishes have been afforded the luxury of diversification in body form not possible for terrestrial animals. A deep, pancake-thin body is not uncommon, and an eel-like form has been independently developed in many phyletic lines. Trunkfishes are enclosed in a boxlike casque, and some deep-sea fishes have eyes at the tips of elongate stalks. Long, trailing fins are frequent, and sargass-

sum fishes develop appendages that serve as holdfasts and for concealment.

**Food habits.** Most fishes are more or less carnivorous and predatory, but there is a wide diversity in food habits. Many fish have numerous slender gill rakers with which they strain microorganisms from the water; others have massive teeth and strongly muscled jaws to aid in crushing mollusks or crustaceans. Browsers, nibblers, and grazers employ specially adapted teeth and jaws to scrape vegetation or small attached animals. Some wrasses pluck parasites from larger fishes, and lampreys parasitize other fishes.

**Reproductive habits.** Reproductive habits are no less varied than feeding behavior. Most fishes are oviparous and scatter their eggs, but nest building and parental care assume a broad spectrum—from a prepared pile of pebbles, through a grassy spherical retreat, to oral incubation or development of a marsupial pouch on the underside of the male pipefish. Viviparity and ovoviviparity have originated along independent lines. Enormously complicated modifications of the anal fin have been evolved to effect insemination of some species in which the young are born alive.

**Economics.** Fishes play an important role in the lives of most people. Per capita consumption of fishery products approximates 10 lb annually in the United States, but that figure is vastly increased in maritime nations or in areas in which other high-protein foods are at a premium. Fishing is a way of life in most primitive cultures, and ranks high among recreational activities in highly civilized peoples. Maintenance and care of home aquarium fishes provide an avocation probably numbering in the millions. An occasional swimmer is killed by sharks; many more die from ciguatera contracted from eating poisonous fish flesh, and venomous fishes take a limited toll in human life and suffering. See CHORDATA; CRANIATA; VERTEBRATA. [R.M.B.]

## Pistachio

A small tree (*Pistacia vera*) and its fruit. The fruit is popularly known as a nut but is botanically classified as a drupe.



Pistachio twig with leaves and fruit.

Pistachios are native to Asia Minor and are adapted to semiarid conditions in the warm temperature climatic zones of the world. Commercial production occurs principally in Turkey, Syria, Iran, Afghanistan, and Italy. A few pistachios are grown in California. In Asia Minor the seeds have long been valued as food for man. Annual importations into the United States range from 2000 to 5000 tons. The nuts are eaten salted and roasted and in ice cream and bakery goods. See NUT CROP CULTURE. [E.F.S.]

## Pitch

That psychological property of sound characterized by highness or lowness. Pitch varies most directly with the frequency of sound waves, and pitch discriminations can be made throughout the frequency range of normal hearing, from about 16 to 20,000 cycles per second (cps).

Within the range of frequencies used in the musical scale, up to about 5000 cps, pitch perceptions are characterized by a periodicity related to the octave arrangement of the scale. A tone that is doubled in frequency appears to be the same musical note one octave higher. Thus two C notes sound more alike than an adjacent C note and D note in the same octave. Some few individuals possess so-called absolute pitch, the ability to judge accurately the pitch level of a musical tone without reference to other tones.

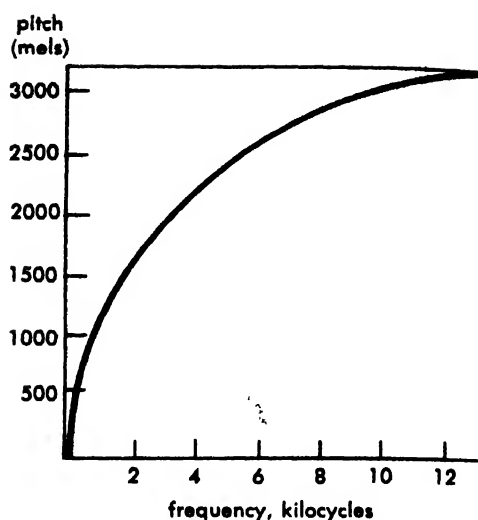
High-frequency sounds sometimes produce a perceived pitch displaced from their characteristic level to the low pitch range. Such a displaced low pitch is not masked by low-frequency tones which would mask a similar pitch produced by a tone of correspondingly low frequency.

A numerical scale of pitch has been constructed by the method of fractionation, selecting pitches that appear to be half as high as reference tones, and the method of bisection, selecting pitches that appear to fall halfway between two reference tones. The pitch unit was named the mel, and a value of 1000 mels was arbitrarily assigned to a tone of 1000 cps.

The pitch of complex sounds may depend on various factors. A musical tone composed of a series of harmonics, such as, 100, 200, 300, 400 cps, is perceived as having a pitch corresponding to the fundamental, 100 cps. This same pitch is perceived even when the fundamental frequency is filtered out of the stimulus. The pitch apparently is determined by the fact that the wave form recurs 100 times per second, even though there is no sound energy at 100 cps.

The pitch of a difference tone corresponds to the difference in frequency of two pure tones presented simultaneously, while the summation tone has a pitch corresponding to the sum of their frequencies.

When a noise stimulus is interrupted from 40 to 200 times per second, observers hear a pitch corresponding to the frequency of interruption.

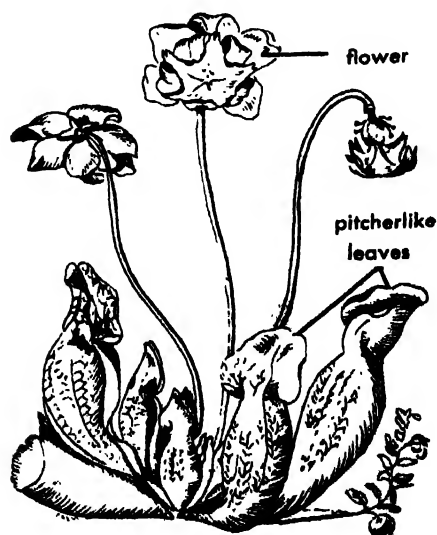


Pitch level of sounds of different frequency as measured in mels. A tone of 1000 cps at an intensity level of 40 decibels above absolute threshold is said to have a pitch level of 1000 mels. (After S. S. Stevens and J. Volkman, *The relation of pitch to frequency*, *Am. J. Psychol.*, 53(3):329-353, 1940)

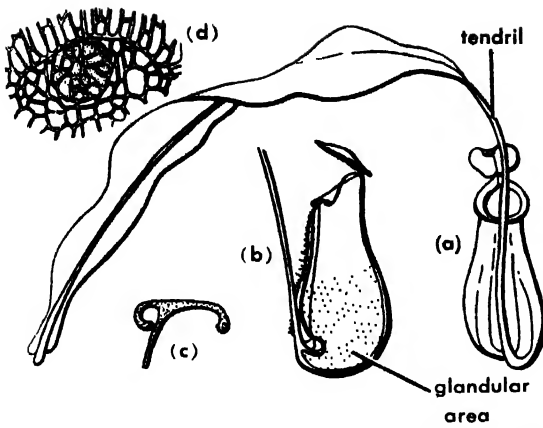
Binaural pitch can be produced by introducing a random noise stimulus to the two ears separately with a constant difference in phase at a certain frequency level (200-1600 cps) within the noise. Observers hear pitchlike sounds corresponding to the level of the phase shift. See DEAFNESS; HEARING. MUSICAL ACOUSTICS. [K.U.S.]

## Pitcher plant

Any member of the families Sarraceniaceae and Nepenthaceae. In these insectivorous plants, the leaves form deep cups or pitchers in which water collects. Visiting insects, falling into this water, are drowned and digested by the action of enzymes secreted by cells located in the walls



*Sarracenia purpurea*, a pitcher plant in flower.



*Nepenthes*. (a) Complete leaf with its pitcher. (b) Vertical section through a pitcher. (c) Section through the margin of a pitcher. (d) Single gland from the lower part of a pitcher. (From R. D. Gibbs, *Botany: An Evolutionary Approach*, Blakiston, 1950)

of the pitcherlike structures of these plants. The Sarraceniaceae are divided into 3 genera; *Sarracenia* in eastern North America, *Darlingtonia* in northern California and southern Oregon, and *Heliamphora* endemic on high mountains in the northern part of South America. The *Nepenthes* family has only one genus, *Nepenthes*, which occurs in the Old World tropics from China to Australia, chiefly in Borneo. Often these plants climb by tendrils (prolongations of the midrib of the leaf). The end of a tendrill may develop into a pitcher, which captures and digests insects. See INSECTIVOROUS PLANTS; SARRACENIALES. [P.D.S.]

### Pitchstone

A natural glass with dull or pitchy luster and generally brown, green, or gray color. It is extremely rich in microscopic, embryonic crystal growths (crystallites) which may cause its dull appearance. The water content of pitchstone is high and generally ranges from 4 to 10% by weight. Only a small proportion of this is primary; most is believed to have been absorbed from the surrounding regions after the glass developed. Pitchstone is formed by rapid cooling of molten rock material (lava or magma) and occurs most commonly as small dikes or as marginal portions of larger dikes. See IGNEOUS ROCKS; VOLCANIC GLASS. [C.A.CA.]

### Pith

The central zone of tissue of an axis in which the vascular tissue is arranged as a hollow cylinder. Pith is present in most stems and in some roots (Fig. 1). Stems without pith rarely occur in angiosperms but are characteristic of psilopsids, lycopsids, *Sphenophyllum*, and some ferns. Roots of some ferns, many monocotyledons, and some dicotyledons include a pith although most roots have xylem tissue in the center. The pith may be present or absent in the same axis, depending upon

size or vigor, the larger segments commonly containing pith.

Pith is composed usually of parenchyma cells often arranged in longitudinal files. This arrangement results from predominantly transverse division of pith mother-cells near the apical meristem. The peripheral region consists of small cells with thick walls and remains alive longer than the central region. When the peripheral region is fairly well defined it is called the medullary sheath or the perimedullary zone. The walls of pith cells may thicken in age, and may become hard or remain soft. In some axes the pith may be composed principally of sclereids. In many fern stems, the inner pith is sclerenchymatous, whereas the outer is parenchymatous. In stems of some dicotyledons, plates or nests of sclerenchyma may be interspersed with the parenchyma. Such a pith is called diaphragmed pith (Fig. 2). If the parenchyma collapses or is torn during development, the scleren-

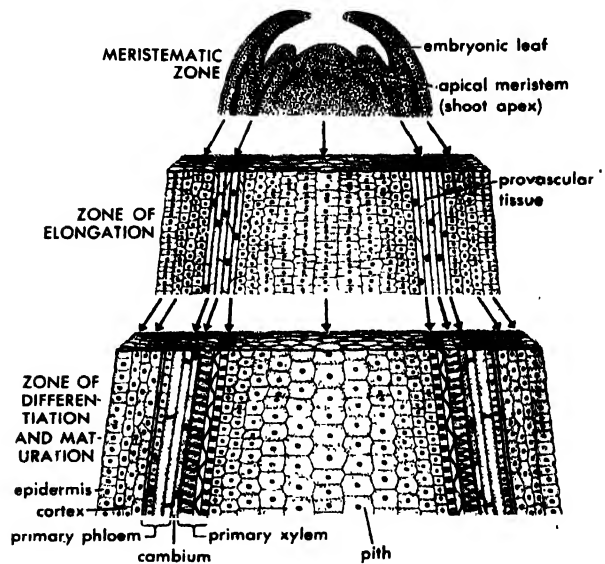


Fig. 1. Zones in the terminal part of a stem. (From C. L. Wilson and W. E. Loomis, *Botany*, rev. ed., Dryden, 1957)

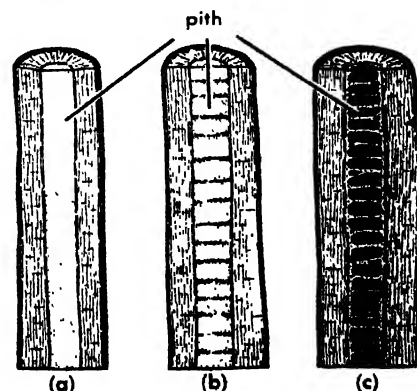


Fig. 2. Types of pith in branchlets. (a) Continuous. (b) Diaphragmed. (c) Chambered. (From E. L. Core, *Plant Taxonomy*, Prentice-Hall, 1955)

chyma plates (diaphragms) alternate with hollow zones. Such a pith is said to be chambered. In many stems the entire pith becomes hollow except at the nodes; the nodal diaphragms are sclerenchyma or parenchyma and, in monocotyledons, may contain vascular bundles.

**Shape of pith.** The shape of the pith in stems of lower vascular plants and in roots of various plants is nearly cylindrical. In stems of higher vascular plants the pith is more or less angled or stellate in cross section. The shape is often characteristic of the plant groups, since it depends on phyllotaxy. In oaks, for example, the pith is 5-angled and in alder, 3-angled. In stems with cylinders of vascular bundles, the panels of ground tissue between bundles often are called medullary rays or pith rays. In stems in which the vascular bundles occur in a more complex arrangement than a simple cylinder, the limit of the pith is indefinite and when a major cylinder of vascular bundles can be distinguished, the internal bundles are called medullary bundles.

**Contents of pith.** Ergastic materials often are stored in some or all cells of the pith. Secretory cells, or secretory canals, or laticifers may be present. In most stems with considerable secondary growth, the pith dies with the formation of heartwood, although the perimedullary zone may remain alive. In other stems the pith may consist partly or largely of dead cells by the end of the first year. See ANGIOSPERMAE; DICOTYLEDONEAE; FILICALES; LEAF (BOTANY); LYCOPODIALES; MONOCOTYLEDONEAE; PARENCHYMA; PERICYCLE; PSILOPSIDA; ROOT (BOTANY); SCLERENCHYMA; SECRETORY STRUCTURES, PLANT; STELE; STEM (BOTANY). [H.W.BL.]

## Pitot tube

An instrument, also called an impact tube, that measures the stagnation pressure of a flowing fluid. Stagnation (also called impact or total) pressure is the pressure that would be obtained if the fluid were brought to rest isentropically. When measuring total pressure, the system consists of a primary sensing element mounted on a suitable support, pressure connecting lines, and a pressure-indicating device. Normally, the connecting lines and indicating devices are considered secondary elements and are not treated as part of the pitot tube. See MANOMETER.

**Application.** The pitot tube is used primarily to obtain fluid velocity, total energy as measured being composed of impact or stagnation pressure  $P_2$ , static pressure  $P_1$  and velocity  $v_1$  of the fluid of density  $\rho$ . Then for incompressible (low-speed) flow

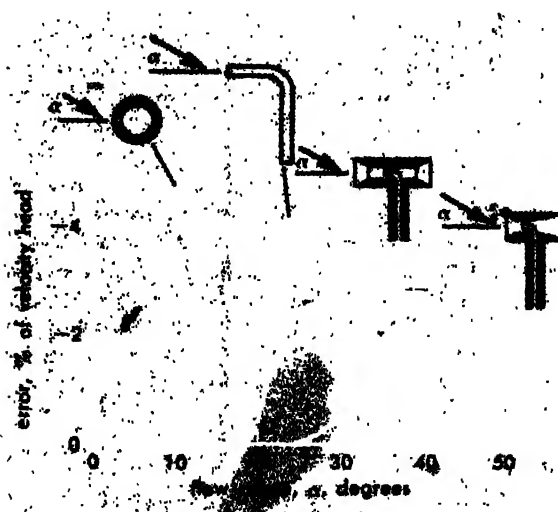
$$\frac{P_2}{\rho} = \frac{P_1}{\rho} + \frac{v_1^2}{2}$$

For incompressible flow, the total energy is expressed by Bernoulli's theorem (see BERNOULLI'S THEOREM). For compressible flow, total energy can

be expressed in terms of impact pressure, static pressure, and Mach number, which is related to velocity.

Pitot tubes of many shapes and sizes are used in a wide variety of applications. A square-ended circular tube pointing upstream will measure true total pressure at subsonic speeds and will measure true total pressure existing behind a normal shock wave across its nose at supersonic speeds.

**Accuracy.** Depending on design, tubes can be made insensitive to flow misalignment up to 45°, as illustrated. Another error arises when a pitot tube is in a total pressure gradient. The effective center of the tube is then displaced from the geometric center toward the region of higher total pressure. Other errors will arise when dealing with turbulent or pulsating flow because of the pressure-averaging effect of the tube.



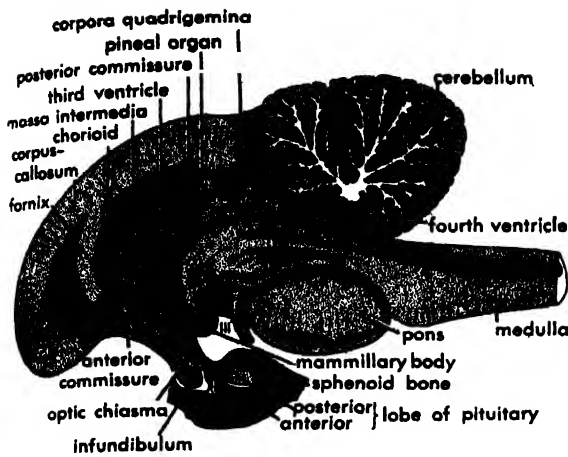
Effect of flow alignment of pitot tube with stream line on accuracy of measurement.

In certain flow regions, conventional pitot-tube response must be corrected to obtain velocity accurately. One such region is that of low Reynolds numbers where the viscous effect of the fluid predominates; another such region is that of high Knudsen number (slip and free molecular flow), which is associated with measurements in a rarefied gas. See AIR-VELOCITY MEASUREMENT; FLOW MEASUREMENT; KNUDSEN NUMBER; REYNOLDS NUMBER. [L.N.K.]

**Bibliography:** R. C. Folsom, Review of the pitot tube, *Trans. ASME*, 78:1447-1460, 1956; L. M. Milne-Thomson, *Theoretical Aerodynamics*, 1952.

## Pituitary gland

The most important single endocrine organ, because its secretions condition many essential metabolic processes. The gland, also known as the hypophysis is present in all vertebrates and is intimately related to the hypothalamus, a portion of the brain from which its main innervation is received. The human gland weighs approximately 0.5 g. and is



Relation of the pituitary gland to the brain and sphenoid bone. (From H. W. Rand, *The Chordates*, Blakiston, 1950)

lodged in the sella turcica, a deep depression in the sphenoid bone. Although considerable anatomical variation is encountered among the vertebrate groups, the gland is roughly divisible into anterior, intermediate, and posterior lobes.

Tissue from two sources enters into the embryonic differentiation of the hypophysis: the anterior and intermediate lobes develop from Rathke's pouch, an evagination from the roof of the embryonic mouth; whereas the posterior lobe originates as an outgrowth of neural ectoderm from the floor of the embryonic brain. The anterior and intermediate lobes eventually lose all connection with the oral epithelium, but the posterior lobe remains permanently connected with the floor of the third ventricle by means of a delicate pituitary stalk. Large numbers of unmyelinated nerve fibers from certain hypothalamic nuclei course through the stalk and terminate in the substance of the posterior lobe.

### ANATOMY

The pituitary is attached to the underside of the forebrain by a more or less hollow stalk of brain tissue, which is generally longer in higher vertebrate classes. Except in a few fishes, the gland rests snugly in a depression (sella turcica) in the basisphenoid region of the skull. In most vertebrates it consists of four fundamental parts: pars glandularis (pars distalis, adenohypophysis, anterior lobe, pars anterior, though often posterior in position); pars intermedia (intermediate lobe); pars neuralis (pars posterior, pars nervosa); and pars tuberalis (lateral lobe).

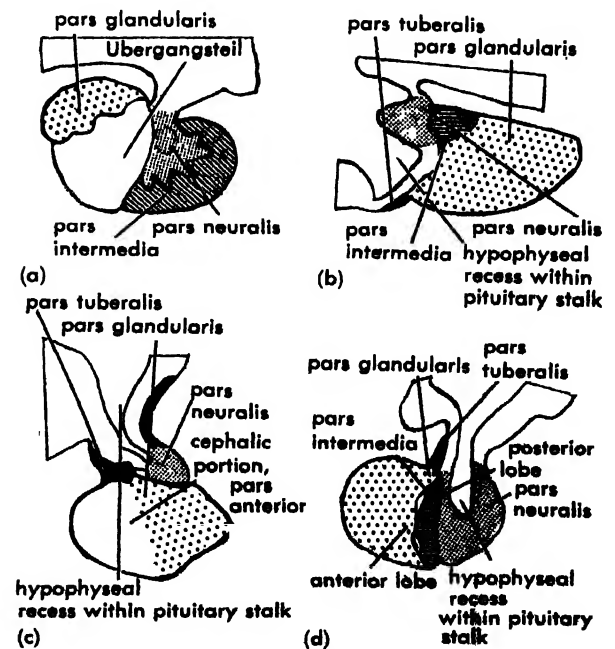
A considerable variation takes place within the vertebrate classes with respect to the orientation of the four subdivisions, and efforts to homologize certain parts, especially in fishes, with parts in higher vertebrates are continuing. The glandularis is usually the most prominent region, but the neuralis becomes increasingly prominent in higher forms. Since names of the parts in lower vertebrates

have been borrowed from mammalian anatomy, the mammalian gland will be described first.

**Mammals.** Mammals usually exhibit a prominent anterior lobe (pars glandularis) and posterior lobe (pars intermedia and pars neuralis) separated in most species, by a lumen of varying proportions. A pars tuberalis, continuous with the glandularis, entwines around the pituitary stalk. Precise orientation of the four components varies. The glandularis and neuralis are prominent. The neuralis often contains a recess of the third ventricle. Between the glandularis and the neuralis lies the intermedia, occupying the posterior wall of the residual lumen when the latter is present. The intermedia is relatively small, and because its limits are not always sharply defined, it may be indistinguishable as a morphologic entity. It is reduced or absent in adult manatees, cetaceans, armadillos, and Indian elephants. The tuberalis is absent in sloths, greatly reduced in manatees, and questionably identified in anteaters.

**Birds.** In birds the gland consists predominantly of a glandularis exhibiting two distinct regions, the more caudal of which resembles the mammalian glandularis. A relatively small neuralis is separated from the glandularis by a distinct connective tissue sheath. No pars intermedia or residual lumen occurs. There is a well-developed tuberalis.

**Reptiles.** In reptiles the pars anterior is an elongated ventral structure attached at its posterior pole only (sauropsids), or at both ends only (chelonians). The intermedia may constitute the major part of the gland (some lizards, and chelonians), or it may be relatively small. Anteriorly, the intermedia is intimately attached to the neuralis, into



Pituitary glands, sagittal sections (cephalic end to left). (a) Teleost. (b) Urodele. (c) Bird. (d) Man.



which the stalk cavity often sends diverticula. No tuberalis has been described in adult snakes and some lizards, but projections of the intermedia in turtles and alligators have been so designated. Apparently, tuberales are present in the embryos of all. The lumen persists in some, disappears in others.

**Amphibians.** In amphibians the pituitary is flattened against the under surface of the mesencephalon. The recess within the short stalk is broadly open to the gland. As a result the neuralis, a plicated, crescent-shaped lobe, is exhibited in its primitive condition as a specialized region of the brain floor. The glandularis constitutes the caudal, major part of the gland. At its anterodorsal boundary a wedge of basophilic cells, usually identified as the intermedia, intervenes between the glandularis and the neuralis. There is no lumen. The tuberales of young animals consist of relatively large, paired, club-shaped appendages, occasionally detached, and extending forward on the under surface of the brain. They vary from  $\frac{1}{4}$  as large as the intermedia to five times its size.

**Fishes.** In bony fishes the gland varies greatly. The glandularis is relatively small; but in most fishes (except lungfishes) a prominent *Übergangsteil* of characteristic morphology and doubtful homology lies between the glandularis and the intermedia. The latter may be deeply penetrated by processes of the neuralis. Although a tuberalis has been described in numerous fishes, the homology of the parts so named is questionable. A ventral lobe hanging by a stalk from the intermedia of selachians has been so designated. A lumen is absent in teleosts and present in selachians. In the primitive *Polypterus*, a lumen within the glandularis opens to the oral cavity. The gland in lungfishes strikingly resembles that of amphibians. However, a secondary lumen sometimes separates intermedia and glandularis, and the latter is occasionally anterior in position.

In cyclostomes the neuralis is scarcely distinguishable as an inconspicuous thickening of the brain floor, under and against which lies the intermedia. The glandularis and *Übergangsteil* lie anteriorly, in tandem with the intermedia. No pars tuberalis or lumen occurs. In hagfishes the embryonic components remain spatially separate, failing to produce a discrete gland. [C.C.K.]

#### HISTOLOGY

The pituitary gland consists of two embryologically and histologically distinct divisions, the glandular division, or adenohypophysis, and the neural division, or neurohypophysis.

In most vertebrates above cyclostomes and fishes the adenohypophysis consists of three distinct regions, the anterior lobe or pars distalis, the intermediate lobe or pars intermedia, and the pars tuberalis. In cyclostomes and fishes the adenohypophysis is partially divided by connective tissue septa into three histologically distinct zones, but the homologies between these zones and the partes

anterior, intermedia, and tuberalis of higher vertebrates have not been established.

**Anterior lobe.** The anterior lobe is the largest subdivision of the adenohypophysis. It consists of irregular cords and masses of glandular cells supported by a network of connective tissue and separated by wide sinusoids. Two main cell types, chromophobes and chromophiles, are distinguished. The chromophobes have few, if any, granules; and their cytoplasm stains poorly with the conventional stains. The chromophiles contain numerous granules and are classified on the basis of stain reactions of these granules into acidophiles (alpha cells) and basophiles (beta cells), respectively. While both types of granule are amphoteric, the acidophiles stain intensely with acid dyes at low pH and only moderately with basic dyes at high pH, while basophiles stain intensely with basic dyes at high pH and poorly with acid dyes at low pH. Two different types of acidophiles and three different types of basophiles have been distinguished in many of the higher vertebrates by such criteria as staining reactions, shape, location, and response to endocrine disturbances. The acidophiles are known as fuchsinophiles and orangeophiles, respectively. The fuchsinophiles appear to be largely restricted to the rostral zone, while the orangeophiles are distributed throughout the lobe. Differential hormone assays have led to the tentative conclusion that the fuchsinophiles elaborate prolactin and the orangeophiles, the growth hormone. Since all basophiles stain with periodic acid-Schiff reagent (PAS), it is generally felt that they elaborate the glycoprotein hormones, the gonadotropins and thyrotrophin. Two types of gonadotrophs or delta cells and a thyrotroph or beta cell have been described. However, some doubt exists as to whether there are several different types of acidophiles and basophiles or whether the tinctorial differences are due to variations in granule content.

In cyclostomes and fishes, chromophobic and chromophilic cells have been described, but they differ somewhat from those of higher vertebrates in distribution and staining characteristics.

Mitoses are rarely seen in the anterior pituitary. However, various intermediate stages have been described between chromophobes and acidophiles on the one hand and basophiles on the other. Chromophobes are generally looked upon as reserve cells which give rise to chromophiles, the actively secreting cells. In the rat, two types of chromophobes have been described. While similar differences have not been observed in other species, it seems to be generally inferred that a particular chromophobe gives rise to one type of chromophile only.

**Blood supply.** The sinusoids of the anterior lobe are lined by a continuous layer of reticulo-endothelial cells. Peri- and intersinusoidal spaces surround sinusoids and connect sinusoids, respectively. Thin basement membranes applied to the gland cells and to the sinusoidal endothelium line these spaces. The spaces contain granules and small



segments of granule-containing cytoplasm. The anterior lobe contains few, if any, nerve fibers.

**Intermediate lobe.** The intermediate lobe consists of a sheet of chromophobic or faintly basophilic cells lying between the anterior and neural lobes. Particularly in the embryo and young animal, it may be separated from the anterior lobe by an extensive cleft, the hypophyseal cleft, the residuum of Rathke's pouch. In other instances the cleft is either missing or represented by isolated, cystlike structures. Both cysts and clefts may be partially lined with ciliated epithelium and may contain colloid and degenerating cells. While the pars intermedia is usually completely separated from the neural lobe by a layer of connective tissue, in man and certain other animals the separation is incomplete, and cells from the intermediate lobe often invade the substance of neurohypophysis. In the ox, a cone of predominantly acidophilic cells, the cone of Wulzen, projects from the intermediate lobe into the hypophyseal cleft and may even protrude into the anterior lobe. Its functional significance has not been determined. A somewhat similar formation has been noted in the sheep.

In certain animals, including birds, cetaceans, the elephant, the manatee, the armadillo, and the beaver, there is no histologically demonstrable pars intermedia; while in others, such as man and man-like apes, it is greatly reduced in size.

The pars intermedia is much less vascular than the anterior lobe. It may contain an occasional nerve fiber which can be traced back to the hypothalamic-hypophyseal tract.

**Pars tuberalis.** The pars tuberalis is a thin layer of cells which arises from the dorsal surface of the anterior lobe, extends upward through the diaphragma sella, and forms a collar around the neural stalk. The cells are chromophobic but are smaller than the chromophobes of the anterior lobe. Small colloid-filled follicles may be present. The vascular supply is rich, and occasional nerve fibers reach the pars tuberalis from the hypothalamic-hypophyseal tract. In a few species, including certain snakes and lizards and the two- and three-toed sloths, there is no pars tuberalis.

**Neurohypophysis.** The neurohypophysis includes the pars nervosa of infundibular process (neural lobe) and the infundibulum or neural stalk, which consists of the infundibular stem and the median eminence of the tuber cinereum. In most species the infundibular process of the mature animal is solid, the infundibular recess of the third ventricle ending within the neural stalk. In some species such as the cat, lion, tiger, bear, sloth, anteater, and pig, the hypophyseal recess extends down into the infundibular process.

The neurohypophysis is a highly vascular region. Its distinctive histologic features include modified neuroglia cells known as pituicytes, a rich plexus of unmyelinated nerve fibers, and a variable number of discrete colloid masses known as Herring bodies. Pituicytes have been described in all classes of vertebrates. They possess cytoplasmic processes

which frequently end in close relation to blood vessels. Their cytoplasm contains granules that stain with fat stains. The function of pituicytes is unknown, though it has been suggested that they play a role in the storage or the transfer of the neural lobe hormones, or both. The nerve fibers arise largely from the supraoptic and paraventricular nuclei in the hypothalamus, stream down the neural stalk in a dense bundle, and terminate largely in the neural lobe. A few penetrate to the intermediate lobe and the pars tuberalis. The cells of origin contain a variable amount of a colloidal substance that stains with the Gomori chrome-alum-hematoxylin stain, while the fibers frequently contain beads of Gomori-positive material. The Herring bodies also stain with Gomori stain. It has not been determined whether the Herring bodies lie free in tissue spaces or represent greatly dilated endings of the nerve fibers. It has been postulated that the Gomori-positive material represents either the active neural lobe hormones or a carrier of the hormone.

Nerve cells have been described in the neurohypophysis of a number of animals including man, birds, cetaceans, the dog, the wolf, the polar bear, and the horse. It has been suggested that these are misplaced elements from the supraoptic nucleus.

[F.O.K.]

#### PHYSIOLOGY

Pituitary hormones are protein or polypeptide in nature. Several biologically active preparations have been extracted from the gland and obtained in highly purified forms. While there is no doubt that they can be separated into distinct entities, some workers feel that these purified fractions may actually be only fragments of some larger hormone molecule. The number of hormonal agents extractable from the anterior lobe does exceed the number of principal cell types in that tissue. The exact form of the molecules as the gland secretes them will remain uncertain until it is possible to isolate them from the circulation in quantities sufficient for chemical studies. While homologous hormones obtained from different vertebrate species exert comparable biological actions, some differences have been detected in the chemistry of the hormone molecules.

Hypophysectomy, surgical removal of the pituitary gland, has been accomplished in many vertebrate species. The most profound disturbances may be outlined as follows: (1) dwarfism in the young animal, and some loss of body tissue in adults; (2) atrophy of the testes, ovaries, and sex accessories in the adult or, in the young, a failure of these organs to attain normal function, the operation rendering the animal completely sterile; (3) atrophy of the adrenal cortices and consequent defects resulting from a deficiency of adrenocortical steroids; (4) atrophy of the thyroid, diminished metabolic rate, and other disturbances resulting from a lack of thyroid hormones; (5) disturbances in pregnancy and lactation, if the operation is

performed at the proper time; (6) blanching of the skin in fishes, amphibians, and reptiles due to impairment of the integumentary pigment cells; (7) profound defects in the metabolism of carbohydrates, proteins, and fats; hypophysectomized animals are unusually sensitive to insulin, the blood sugar tends to diminish, and the glycogen stores are depleted rapidly during fasting; the loss of nitrogen from the body indicates excessive protein breakdown, and the catabolism of fats is diminished. See METABOLIC DISORDERS.

The above effects of hypophysectomy are due to the absence of hormones from the anterior and intermediate lobes. Ablation of only the posterior lobe produces much milder defects. Among mammalian species, the only profound effect resulting from postlobectomy is an excessive loss of water through the kidneys, or diabetes insipidus; this condition may be transient or permanent, depending on the species and the manner of performing the operation. There are cogent reasons for believing that certain hypothalamic nuclei, with their fiber tracts, must be considered together with the posterior lobe per se as constituting a functional unit. Since the hormones extractable from the posterior lobe are probably neurosecretions originating in the hypothalamus, it is apparent that postlobectomy may not completely and permanently eliminate these secretions from the system. See HYPOPHYSIS.

**Anterior pituitary.** The anterior pituitary produces at least six principles: five trophic hormones and a growth hormone. The trophic hormones directly control to some degree the functional capacity of another endocrine tissue. There are three gonadotrophins which regulate the endocrine secretions of the gonads, adrenocorticotrophin (ACTH), which conditions the secretion of adrenocortical steroids, and thyrotrophin (TSH), which influences the formation of thyroid hormones. While the growth hormone does promote growth, it exhibits a great variety of other metabolic effects. See HORMONE, ADENOHYPOPHYSAL.

The three gonadotrophins are follicle-stimulating hormone (FSH), luteinizing hormone (LH), and luteotrophin (lactogenic hormone). The main action of FSH in the female is to promote growth of the ovarian follicle up to the point of ovulation; in the male it stimulates the seminiferous tubules and maintains production of sperms. LH cooperates with FSH during the final stages of follicular development, the two hormones promoting the secretion of estrogen and causing ovulation; then LH and luteotrophin stimulate the formation of luteal tissue and thus promote the secretion of progesterone. In the male, LH stimulates the interstitial tissue of the testis and thereby increases the output of androgen. It is clear that FSH and LH have equally important roles in regulating both male and female sexual functions. Both hormones are glycoproteins and have been prepared in highly purified forms. See OVARY; TESTIS.

**Luteotrophin.** This hormone assists in the maintenance of the corpora lutea once they are formed. Perhaps the main action of this gonadotrophin in mammals is to initiate and maintain lactation by acting on the fully developed mammary glands. It stimulates the secretion of crop milk in pigeons and initiates broodiness in certain species. Luteotrophin is the first pituitary hormone to be isolated in pure form; it is a protein having a molecular weight of around 30,000. See LACTATION.

**Chorionic gonadotrophins.** These gonadotrophins, having anterior pituitarylike effects, are secreted by the placenta. Two such substances have been isolated: they are serum gonadotrophin (PMSG) of pregnant mares and chorionic gonadotrophin (HCG) of humans. PMSG elicits actions which are comparable to a mixture of pituitary gonadotrophins, but it has not been possible to separate it into fractions like those of the pituitary. The main action of HCG is on the corpus luteum or on the interstitial cells of the testis. While both PMSG and HCG are glycoproteins, they appear to be chemically different from FSH, LH, and luteotrophin. Gonadotrophins of placental origin are abundant in the blood and urine of pregnant women, the blood of pregnant mares, and in the urine of patients suffering from certain genital tumors.

**Adrenocorticotrophin.** When adrenocorticotrophin stimulates the adrenal cortex, it accelerates the secretion of adrenal steroids. It restores to normal the atrophic cortices of hypophysectomized animals. ACTH preparations can be bioassayed on the basis of their ability to maintain the weights of adrenals of hypophysectomized rats or their capacity to cause depletion of ascorbic acid and cholesterol from the adrenals of hypophysectomized subjects. It should be stressed that ACTH peptides exert a great variety of extra-adrenal actions. There is increasing evidence that all the pituitary trophic hormones produce metabolic alterations by acting on tissues other than their target glands.

ACTH has been isolated as a protein with a molecular weight of around 20,000; others find that ACTH activity resides in much smaller molecules (peptides) which are several times more potent by weight than the original protein ACTH. It is possible that in the gland the peptides are adsorbed on or combined with cellular protein. See ADRENAL GLAND.

**Thyrotrophin.** The normal structure and function of the thyroid gland in hypophysectomized animals is maintained by thyrotrophin. TSH is a small protein having a molecular weight of approximately 10,000; it has been obtained in highly active form but has not yet been purified. A dynamic balance appears to be established between the secretion of TSH by the pituitary and the titers of thyroid hormone in the blood, but other controls seem to exist also. See THYROID GLAND.

**Growth hormone.** Growth hormone is a simple protein having a molecular weight of about 45,000.

and an isoelectric point at pH 6.85. It consists of two chains, one having alanine and the other phenylalanine as the *N*-terminal amino acid, and contains some 369 amino acid residues. The hormone may be assayed by determining its ability to promote widening of the tibial epiphysis in hypophysectomized rats or by techniques involving the capacity of the hormone to encourage nitrogen retention.

Hypophysectomy of young animals retards general body growth; the administration of growth hormone restores the growth rate and alleviates some of the other metabolic disturbances. The hormone exerts a great variety of effects other than just the promotion of general body growth. It produces the typical signs of pancreatic diabetes and causes various changes in carbohydrate metabolism; it causes the retention of nitrogen (as protein) in the tissues; it accelerates the mobilization and oxidation of depot fat; and in some species at least, it is essential for normal secretion of milk. See CARBOHYDRATE METABOLISM.

Certain instances of human dwarfism result from a deficiency of growth hormone during childhood. Hypersecretion of the anterior lobe in the young individual produces a condition of excessive growth called gigantism; if the onset is after the epiphyses of the long bones have closed, acromegaly results.

*Melanocyte-stimulating hormones.* Melanocyte-stimulating hormones (intermedin) are found in the hypophyses of all vertebrates from selachians and teleosts to mammals. These substances are produced by the intermediate lobe or, in those species where the intermediate lobe is anatomically absent, by equivalent areas of the anterior lobe. In fishes, amphibians, and reptiles, MSH is responsible for the color changes which occur when these animals are kept in darkness.

Two polypeptides,  $\alpha$ -MSH and  $\beta$ -MSH, have been isolated from pituitary tissue and chemically characterized. Both produce darkening of the melanocytes in man and certain lower vertebrates. Both substances have molecular weights of less than 4000 and contain approximately 15 different amino acids. Purified ACTH peptides exert slight but perceptible MSH activity, but the reverse is not true; that is, purified MSH possesses no intrinsic adrenocorticotrophic activity.

Whether MSH has any function in birds and mammals has not been conclusively demonstrated. There is some evidence that in man the adrenal cortical steroids inhibit the release of MSH from the pituitary. Some workers believe that in clinical conditions where adrenocortical secretions are decreased or absent the pituitary releases excessive MSH, which causes melanogenesis. In a variety of clinical conditions characterized by intensive melanosis, there are increased amounts of MSH in the blood.

**Posterior lobe.** Posterior lobe extracts exert profound effects upon water-salt balance, the cardiovascular system, respiratory activity, smooth mus-

cle of the uterus and parts of the gastrointestinal tract, and myoepithelial cells ("basket cells") of the mammary glands. O. Kamm et al. in 1928 succeeded in separating fractions which gave principally oxytocic and vasopressor effects. Antidiuretic activity was possessed by the vasopressor fraction. The oxytocic fraction has been used routinely in obstetrical practice to induce labor, the gravid uterus at or near term being very sensitive to the action of this substance. See HORMONE, NEUROHYPOPHYSEAL.

V. Du Vigneaud and associates have obtained purified preparations of oxytocin and vasopressin and established their structural formulas. Both are basic peptides containing eight amino acid residues, six of the amino acids being identical in kind and position in the two hormones. Vasopressin possesses all the antidiuretic activity. Biologically there is some overlap: vasopressin possesses slight intrinsic oxytocic activity, although oxytocin has no vasopressor or antidiuretic action. Oxytocin was the first polypeptide hormone to be synthesized.

In addition to the polypeptides, a protein having a molecular weight of around 30,000 has been obtained from posterior lobe tissue. This substance appears to be pure by all the usual criteria and possesses oxytocic, vasopressor, and antidiuretic activities. Some regard this protein as the true hormone which combines the properties of the fractions. Final conclusions cannot yet be drawn.

In some species the ejection of milk during suckling or milking depends upon the reflexive release of oxytocin from the posterior lobe. Whether oxytocin normally plays any essential role in parturition has not been clearly established.

Vasopressin regulates the amount of water filtered by the kidneys. When most organisms are deprived of water, increased amounts of vasopressin are released from the posterior lobe and water loss is inhibited. After removal of the posterior lobe, without damage to the hypothalamus, the animal loses a large amount of dilute urine (polyuria). As a result, there is increased thirst and a high intake of water (polydipsia). Diabetes insipidus, the clinical counterpart of this condition, has long been recognized. It is not known whether the posterior lobe hormones normally perform any significant physiological role in the regulation of blood pressure. See URINARY BLADDER DISORDERS; URINARY SYSTEM.

Most evidence indicates that the posterior lobe hormones are secreted by special cells in certain hypothalamic nuclei, pass along the axons within the pituitary stalk, and are discharged in the posterior lobe per se. Thus the posterior lobe may be mainly a depot for the storage and release of hormones and not a true endocrine gland, since the hormones are synthesized elsewhere. [C.D.T.]

*Bibliography:* H. Heller, *The Neurohypophysis*, 1957; G. Pincus and K. V. Thimann, *The Hormones*, vol. 3, 1955.

## Pituitary gland disorders

Any pathology of the pituitary gland, the master regulatory organ of the endocrine system. Many hormones which directly influence tissue metabolism are elaborated by the pituitary, but more important, certain of these hormones exert a control on other endocrine, or target glands, such as the thyroid, adrenals, and gonads. The pituitary is actually a double gland, the anterior portion arising from an evagination of the roof of the mouth and the posterior lobe developing as a downgrowth of modified neural tissue from the hypothalamus. The two lobes come together in a bony depression located behind the optic chiasma at the base of the brain. See PITUITARY GLAND.

The most common pituitary disorders can be grouped into two categories, those in which symptoms result from mechanical changes in the gland and those in which hormonal imbalances are produced. Hemorrhages, tumor formation, and inflammatory changes may cause symptoms by virtue of their direct effects. Hormonal dysfunction, although frequently caused by neoplasia, may arise in a grossly unaltered gland, yet have pronounced effects upon other tissues. See HORMONE; TUMOR.

The best-known examples of anterior lobe dysfunction are the cases of gigantism and dwarfism which result from either hyper- or hyposecretion of growth hormones during childhood. In adults, hypersecretion of the growth hormones results in acromegaly, marked by the appearance of enlarged jaws, hands, and feet and other changes.

Hypopituitarism indicates a decreased output of pituitary hormones, and like hypersecretion, may involve one or more or all of the endocrine secretions. Cases of Simmonds' disease follow loss of all secretion and are characterized by a general wasting and secondary insufficiency of the adrenal, thyroid, and gonadal glands, because of lack of tropic stimulation. Adult women are affected twice as frequently as men, and the clinical course may be rapid or may be prolonged over a period of years, depending upon the nature of the pathologic process causing the glandular deficiency.

Deficiency of gonadotropic hormones, those which stimulate either the testes or ovaries, may result in Froehlich's syndrome. This is marked by excessive obesity of the female type, failure of development of secondary sexual characteristics, and sexual dysfunction. In these cases, malfunction of other parts of the pituitary or hypothalamus is probably responsible for the obesity.

The most important, though uncommon, disease of the posterior pituitary is diabetes insipidus. In these patients, a deficiency of the antidiuretic hormone permits water to pass through the renal tubules without proper reabsorption. As a result there is often a tremendous urinary output, sometimes as much as 20-30 quarts, or more, of urine a day.

Most of the pituitary disorders mentioned, as well as those which are less common or complex in

nature, result from injury to, or tumors of, pituitary tissue. Injury may follow trauma, inflammation, vascular damage, or the development of cysts. Frequently the cause of hypersecretion is the overactivity of a benign or malignant tumor; both, however, may be asymptomatic. [E.G.ST.]

## pK

The logarithm (to the base 10) of the reciprocal of the equilibrium constant for a specified reaction under specified conditions (for example, solvent and temperature).  $pK$  values are often more convenient to tabulate and use than the equilibrium constants themselves. The value of  $K$  for the dissociation of the  $HSO_4^-$  ion in aqueous solution at  $25^\circ C$  is 0.0102 mole/liter. The logarithm is  $0.0086 - 2 = -1.9914$ .  $pK$  is therefore  $+1.991$ . The choice of algebraic sign, although arbitrary, results in positive values for most dissociation constants applicable to aqueous solutions. The concept of  $pK$  is especially valuable in the study of solutions. See EQUILIBRIUM, CHEMICAL; EQUILIBRIUM, IONIC;  $pH$ . [T.F.Y.]

## Placentation

The intimate association or fusion of a tissue or organ of the embryonic stage of an animal to its parent for physiological exchange designed to promote the growth and development of the young. It enables the young, retained within the body or tissues of the mother, to respire; acquire nourishment, and eliminate wastes by bringing their blood streams into close association but never into direct connection (Fig. 1). Placentation characterizes the early development of all mammals except the egg-laying duckbill platypus and spiny anteater. It occurs in some species of all other orders of vertebrates except the birds. In fact, in certain sharks and reptiles, it is almost as well developed as in mammals. A few examples are also known among invertebrates (*Peripatus*, certain tunicates, and insects).

**Placental modifications.** With few exceptions, the fetal structures used to establish placental relationships with the mother are modifications of organs present in kindred egg-laying (oviparous) species. In the fishes, sharks and rays, and amphibians these include gill filaments, the tail fin, the pericardium, and the primitive yolk sac (midgut and adjacent ventral body wall). The essential placental modifications of all these are increased surface area and vascularity and intimate contact with a highly vascularized and often secretory area of the mother. In fishes this latter is usually the ovary; in sharks and rays, the uterus; in amphibians, the uterus or the skin. Skin gestation is unique to certain oviparous South American frogs. As the eggs are extruded, they are fertilized by the male and placed on the back as in *Rana* or in a skin pouch on the back as in *Nototrema*. Here they become embedded in highly vascular compartments and receive both oxygen and nourishment from the mother's blood stream. The live-bearing fishes

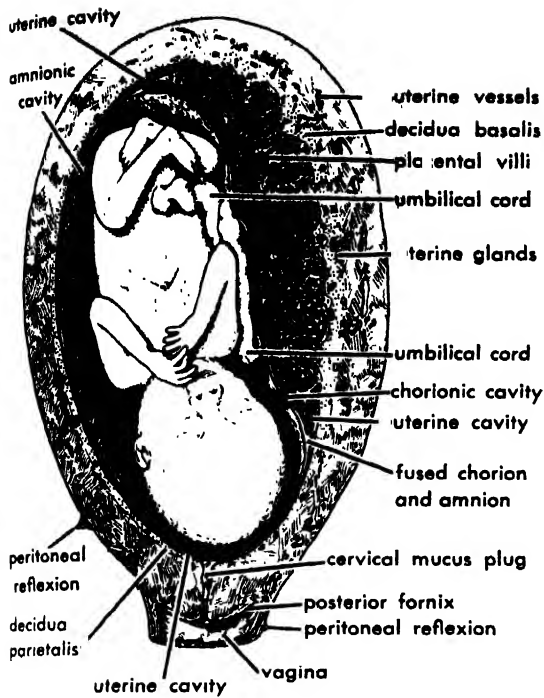


Fig. 1. Pregnant human uterus at 3½ months, split sagittally to show the relation of the fetus, membranes, and placenta.

Goodeidae, have unique vascular rectal processes, trophotaeniae, known in no other group, which establish placental relationships with the ovarian tissue.

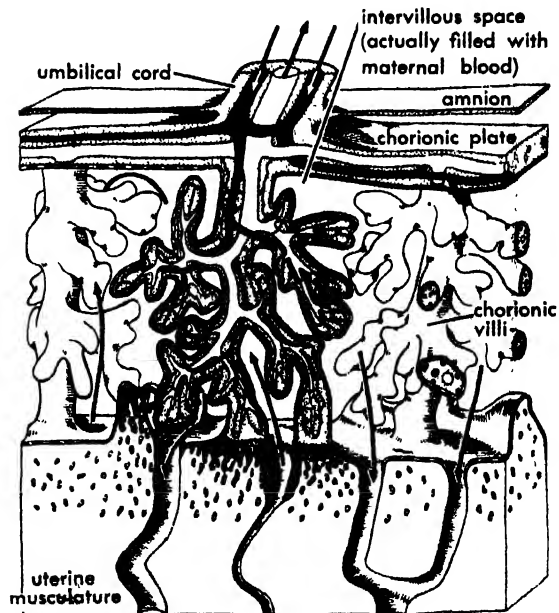
In amniotes (reptiles, birds, and mammals) the extraembryonic membranes utilized in placentation are specializations of the basic membranes found in all the oviparous species. In fact, in the beginnings of placentation, portrayed by ovoviviparous lizards and snakes, the large shell-covered and yolk-laden eggs are simply retained within the uterus until hatched. Here apparently the only placental function is respiration, hence there is no obvious modification of the fetal membranes and only an increased and prolonged hypervascularity of the uterine lining.

In a few lizards and snakes and in all marsupial and eutherian mammals, complete placental function occurs because the eggs are supplied with too little yolk to provide the needs of the embryo for nourishment until birth. Inadequate provision for elimination and storage of nitrogenous waste occurs in most of these animals also. Thus, not only must these embryos interchange respiratory gases with the mother's blood, but they must also absorb nutriment and transfer wastes to be excreted by her kidneys. Full placentation in these forms is provided by anatomical and physiological specialization of three principal extraembryonic membranes, the chorion, yolk sac, and allantois. The amnion probably plays a physiological role also.

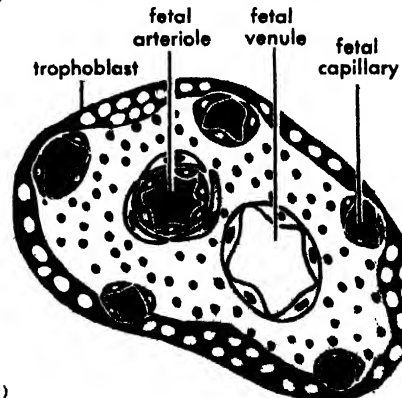
**Physiological exchange.** Efficient interchange depends on close proximity of large areas of fetal tissues to maternal blood and to secretory or absorp-

tive areas. This is provided in mammals by a remarkable regulatory cooperation between the developing outer layer (trophoblast) of the chorion, together with the vascular yolk sac or allantois, both, and the mother's uterine lining (endometrium). In the typical mammalian placenta, which is always formed by the chorion and the allantoic vessels, the fetal and maternal blood streams are as close as a few thousandths of a millimeter from each other (Fig. 2a,b). The surface area is probably several times larger than the body surface of the female. In man this ratio is known to be about 8:1. Not only are the two vascular tissues closely approximated, but in many mammals the arrangement of the vessels is such that in the area of interchange the two blood streams flow in opposite directions. This counterflow principle greatly increases the efficiency of interchange.

Placentas are classified in various ways, but the most meaningful is based on the identity of the layers making up the separation membrane between



(a)



(b)

Fig. 2. (a) Diagram of block removed from center of human placenta. (b) Enlargement of cut end of branch villus.

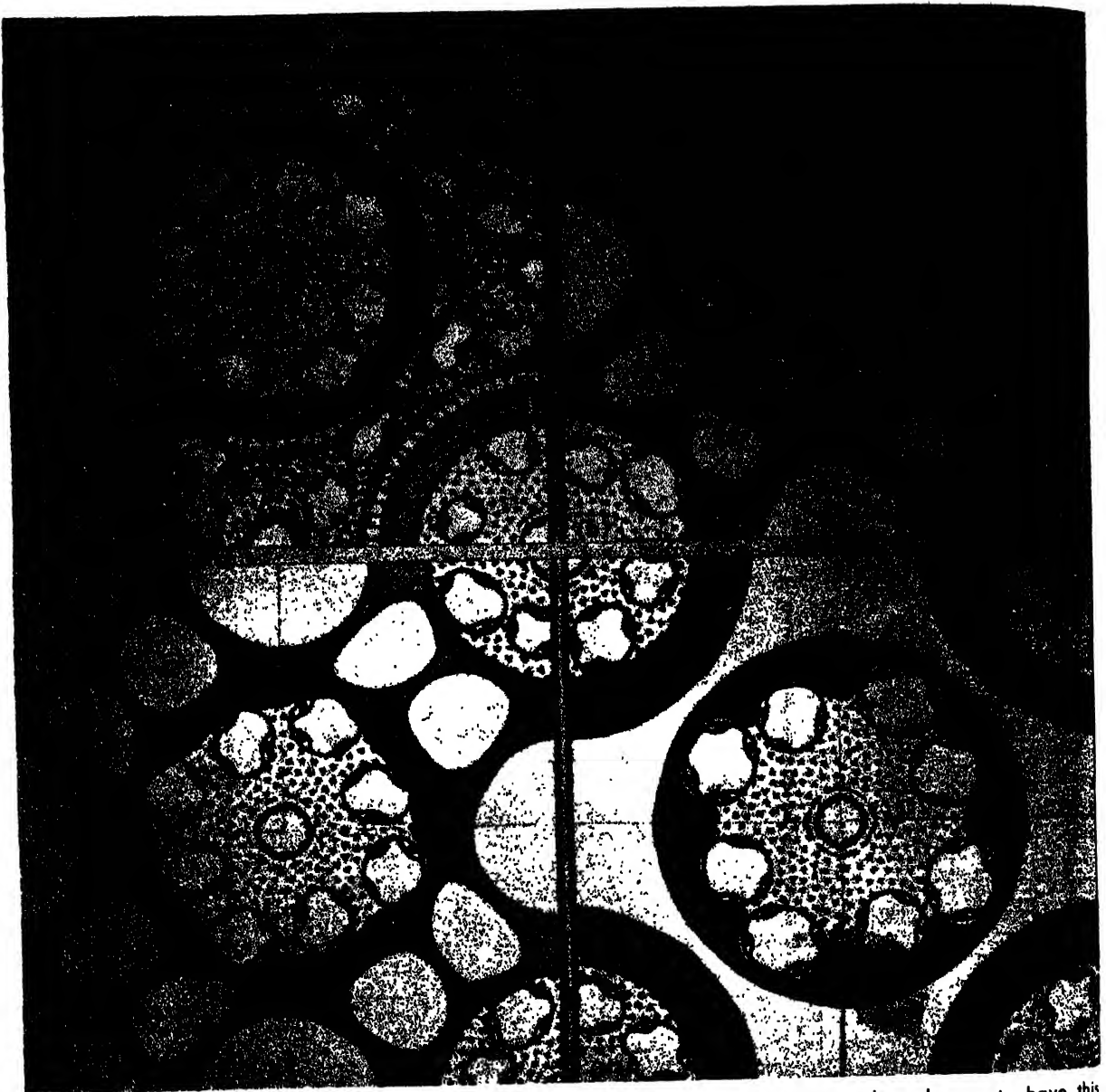


Fig. 3. Schemata of four fundamental types of placentation. (a) Epitheliochorial (villous); examples: hoofed animals, whales, lemurs. (b) Endotheliochorial (labyrinthine); examples: carnivores, bats, elephant, sloth,

beaver (one of only two rodents known to have this type). (c) Hemochorial (labyrinthine); examples: most rodents, many insectivores, tarsiers. (d) Hemochorial (villous); examples: man, great apes, and monkeys.

the two blood streams. On this basis the human placenta is hemochorial, that is, the maternal blood is in direct contact with the chorionic trophoblast. The placenta of the dog is endotheliochorial, in which the maternal blood is separated from the chorion by the maternal capillary endothelium. Other conditions are epitheliochorial, syndesmochorial, hemoendothelial, and endothelioendothelial.

The physiology of interchange through the placental membrane is a fertile field in present-day research. Although simple physical diffusion and osmosis are factors, it is now known that probably of greater importance are active membrane transport mechanisms. This transport through the placental separation membrane is work done as the

result of energy release within the protoplasmic layers of the membrane itself. [H.W.MO.]

**Bibliography:** W. J. Hamilton, J. D. Boyd, and H. W. Mossman, *Human Embryology*, 2d ed., 1952; A. S. Parkes (ed.), *Marshall's Physiology of Reproduction*, vol. 2, 3d ed., 1952.

### Placodermi

A large and varied assemblage of Paleozoic fishes, usually categorized as a zoological class, which appeared only slightly later than the Acanthodii. The Placodermi constitute the second oldest group of primitive jawed vertebrate animals. As implied by their name, which means plate-skinned, the placoderms are characterized by the development of a complex bony armor protecting the head and



front part of the body. Typically, the shield covering the head is separate and is movable upon the body armor by paired ball-and-socket articulations. The oldest known placoderms are of Late Silurian age. During the Devonian they achieved a dominance over all other contemporaneous groups of backboned animals. Their success, however, was short lived and by the close of the Devonian (except for *Cratoseleache* from Mississippian rocks of Belgium) these fishes became extinct. See ACANTHODII.

The placoderms are separable into two major divisions, the Arthrodira and Antiarchi. Within a widely displayed range of shape, size, and anatomical variation, both groups possessed a few common characteristics. The neurocranium formed chiefly of cartilage was a stout structure composed of a short ethmoidal part supporting large olfactory organs, moderate-sized and forward-situated orbital indentations, and a relatively long, robust oticooccipital portion. Various dermal bones of vomerine and parasphenoidal nature often invested the ventral surface of the braincase, with the former sometimes modified as integral parts of the biting mechanisms of the upper jaw. Dorsally and laterally a dermal shield composed of a regular series of median unpaired and paired lateral plates of bone protected the neurocranium, visceral skeleton, and branchial chambers. The visceral skeleton, perhaps through persistence of a primitive soft cartilaginous condition and consequent failure to be preserved in the fossil state, remains most incompletely understood. However, partial endochondral ossifications offer evidence of the mandibular arch in many arthrodires. These paired ossifications are located in the quadrate areas of the upper jaw and the articular and symphyseal regions of the lower. Each of the primary upper- and lower-jaw elements supported single pairs of opposing bones. Adapted for either crushing or scissorlike shearing action, these elements were not deciduous but continued to grow and enlarge during the life of the individual, often displaying enameled denticles and columns of a dentinal-like tissue. Pits and rugosities on the margins of the quadrate and articular ossifications of the jaws suggest ligamentous attachments, very probably with a modified element of the second visceral arch. Thus, for at least some of the placoderms, an amphistylic type of jaw articulation with the braincase may be postulated with a suspensory hyomandibula and restriction of hyoidian gill cleft to its dorsal spiracular part. See ANTIARCHI; ARTHRODIRA.

The thoracic armor, as that of the cephalic shield, is composed of median unpaired and paired lateral elements. These dermal bones are connected by both overlapping sutures and tightly interlocking dentate ones. The buckler is most comparable in position and function to the exoskeletal shoulder girdle of more advanced bony fishes and displays vastly different developments in each of the varied placoderm stocks. Among the earliest as well as the most conservative lines it forms a continuous

shield over the back, sides, and venter of the trunk. In the later, younger forms it may become variously modified and reduced, often with complete losses of connection between the dorsal and ventral parts and even of the typical mobile articulation with the head armor. The body and tail behind the trunk armor is generally fishlike and may be either scaled or naked. The internal axial skeleton is known adequately in only a few forms. In the arthrodire *Coccosteus* the notochord was persistent and extended to the tip of the heterocercal caudal fin. In addition, segmental condensations of bony tissue reveal dorsal neural and ventral hemal arcualia. One or two median dorsal fins were present, supported internally by rays. The hypochordal lobe of the caudal fin is presumed to be small. Of paired fins, pectorals and pelvics were generally present. The pectoral appendages were of variable form. A pair of fixed or movable spines may alone project from the sides of the thoracic armor in the position of these fins. In arthrodires the fixed spines became progressively reduced through geologic time and their function was gradually replaced by proportionately larger and larger normal fins with internal rays articulating on a primary cartilaginous shoulder girdle snugly associated with the dermal trunk armor.

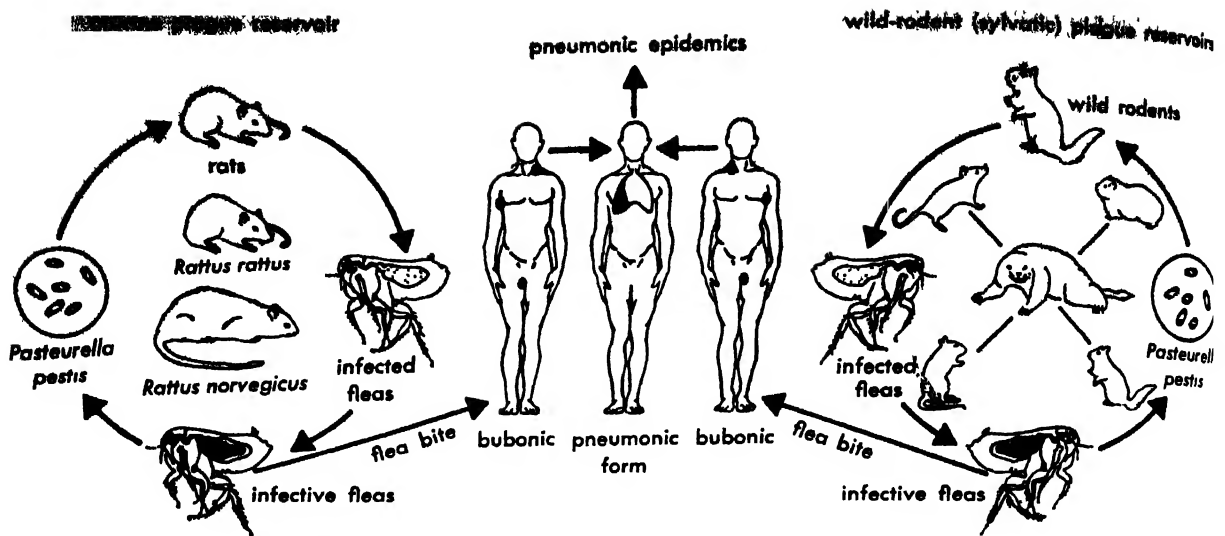
The Placodermi form an extraordinary group of fishes whose remarkably diverse adaptation to life in both fresh and marine waters makes any phyletic interpretations difficult. In the present state of knowledge, they can only be regarded as an early aberrant assemblage of armored backboned animals whose resemblances to acanthodian, chondrichthyan, and early osteichthyan fishes were a heritage from a common, early Paleozoic ancestral stock of jawed vertebrates. See PALAEOSPONDYLOIDEA; STEGOSELEACHII. [D.H.D.]

## Plague

An infectious disease of man and rodents, existing as pneumonic and bubonic plague and caused by the bacillus *Pasteurella pestis*. In wild rodents the disease is known as sylvatic plague. Plague is transmitted from rodent to rodent and from rodent to man by the flea (*Xenopsylla cheopis*). The disease has been known since the third century for its pandemics and epidemics such as the Black Death in 1630 (Milan), in 1665 (London), and in 1721 (Marseilles). Alexander Yersin discovered the cause of the disease in 1894 during a worldwide outbreak in the 1890s.

In the animal body the bacillus is gram-negative, nonmotile, and a short, round coccoid or a large ovoid, safety pin in shape. It is surrounded by a slime layer, or envelope. This aerobic organism grows well, but slowly, on medium containing cystine and blood at 37°C, the time required for cell division being about 4 hours. Glucose, galactose, fructose, but not lactose, are fermented without production of gas. The ability to ferment glycerol is characteristic of certain varieties of *P. pestis* and is an aid in determining the epi-





Epidemiology of plague. (George Williams Hooper Foundation, University of California)

demology of an outbreak of the disease. *P. pestis* fails to hydrolyze urea. Under favorable laboratory conditions it remains viable for months, even years; in rat flea pellets, for months. It is inactivated at 55°C by 0.5% solution of phenol in 15 min and by streptomycin and tetracycline. Of its at least 10 antigens, the envelope antigen and the powerful toxic antigen are significant in immunity and pathogenesis.

Plague may be diagnosed by culture of blood or tissue fluid, obtained from lymph nodes of man or dead rodents or by animal inoculation. Cultures may be quickly identified by a specific bacteriophage or more readily by agglutination test with potent antiserum. After the diagnostic material is rubbed on the freshly shaved abdomen of guinea pigs, they die in 2-6 days; bacilli can be found in blood and spleen films.

In warm climates plague usually is bubonic, so called because of the characteristic swollen lymph nodes, or buboes. Bubonic plague is usually transmitted to man by rat fleas. It may spread rapidly at times among commensal rats and mice, mainly through the rat flea. During rat epizootics these fleas carry the plague bacillus in the midgut and proventriculus from which it is regurgitated during the bite. In cold climates it is more likely to take the much more fatal and contagious pneumonic, or tonsillar, form. Pneumonic plague spreads from patients with primary pneumonic plague or from patients with bubonic plague and secondary pneumonic infection.

Endemic (sylvatic) plague may be maintained by any of 372 burrowing, hibernating rodents, such as field mice, squirrels, prairie dogs, wood rats, spermophiles or marmots, which are widely distributed in the western third of the United States, in large areas of South America, in Central and South Africa, in Iranian Turkistan and Central Asia.

Human infection is rare in the United States and arises from exposure to wild rodents or their

fleas. Plague is often a terrible health problem in cities in India, Burma, and Indonesia, although it is presently in a decline. It still persists there however, in the wild rodents.

After treatment with sulfonamides (particularly sulfadiazine), streptomycin, tetracycline or chloramphenicol, spectacular effects have been observed, even in pneumonic plague which was formerly irremediable.

Perpetual systematic warfare against rodents is fundamental to prevention. The potent rodenticide Warfarin (dicoumarin) is used to free cities of rats and to establish rodent-free belts around towns and villages. In the control of epidemics the first consideration must be elimination of the flea by use of insecticides with residual action (DDT in 5% kaolin or malathion). See AGGLUTINATION REACTION; BACTERIOLOGY, MEDICAL; BACTERIOLOGY; BRUCELLACEAE. [K.F.M.]

*Bibliography:* R. Pollitzer, *Plague*, WHO Monograph 22, 1954.

## Plains

The relatively smooth sections of the continental surfaces, occupied largely by gentle, rather than steep slopes and exhibiting only small local differences in elevation. Because of their smoothness, plains lands, if other conditions are favorable, are especially amenable to agricultural use. The absence of extensive steep-sloped features of great height not only decreases the number of obstacles to human transportation or, in past times, to the migrations of animals, but also permits the free movement of air masses, a fact of great meteorological importance. It is thus not surprising that the majority of the world's major agricultural regions, closely meshed transportation networks, and concentrations of population are found on plains. Nor is it unexpected that extensive plains are areas within which climatic conditions vary but slightly over long distances.

**Distribution and varieties.** Somewhat more than one-third of the earth's land area is occupied by plains. With the exception of ice-sheathed Antarctica, each continent contains at least one major expanse of smooth land in addition to numerous smaller areas. The chief plains of North America, South America, and Eurasia lie in the continental interiors, with extensions reaching to the Atlantic. The most extensive plains of Africa occupy much of the Sahara and reach south into the Congo Basin. Much of Australia is smooth, with only the eastern margin lacking plains terrain. See TERRAIN AREAS, WORLD-WIDE.

Surfaces that approach true flatness, while not rare constitute a minor portion of the world's plains. Most commonly they occur along low-lying coastal margins or the lower sections of major river systems. Some occupy the floors of inland basins where extensive stream deposition has occurred. The majority of plains, however, are distinctly irregular in surface form, as a result of valley-cutting by streams or of irregular erosion and deposition by continental glaciers.

Plains are sometimes designated by the situations in which they occur. In common speech a coastal plain is any strip of smooth land adjacent to the shoreline, though in geology the term is often restricted to such a plain that was formerly a part of the shallow sea bottom. An example is the South Atlantic and Gulf margin of the United States (see COASTAL LANDFORMS; COASTAL PLAIN). Intermontane plains lie between mountain ranges, and basin plains are surrounded by higher and rougher land. Upland plains (sometimes loosely termed plateaus) lie at high elevations, or at least well above neighboring surfaces, while lowland plains are those lying near sea level, or distinctly below adjacent lands.

Plains are also sometimes classified according to the processes that have produced their distinctive surface features. These differences are discussed below.

**Origin.** The existence of plains terrain generally indicates for that area a dominance of the erosional and depositional processes over the forces that deform the crust itself. The most extensive areas of plains, such as those of interior North America or that of northwestern Eurasia, generally represent areas which have experienced nothing more severe than slow, broad warping of the crust over a long period of geological history. Throughout that time the gradational processes have been able to maintain a relatively subdued surface. Certain other areas, including the upland plains of central and south-central Africa and eastern Brazil, have suffered moderate general uplift in late geologic time, but have not yet been subjected to deep valley cutting.

Many plains of lesser extent, however, have been formed in areas where crustal deformation has been intense. Most of these represent depressed sections of the crust which have been partially filled by smooth-surfaced deposits of debris carried

in by streams from the surrounding mountains. Examples of such plains are the Central Valley of California, the Po Plain of northern Italy, the plain of Hungary, the Mesopotamian plain, the Tarim Basin of central Asia, and the Indo-Gangetic plain of northern India and Pakistan.

#### REGIONAL SURFACE CHARACTER

The surface features of plains result mainly from local erosional and depositional activity in relatively recent geologic time. Each of the major gradational agents—running water, glacial ice, and the wind—produces its own characteristic set of features, and any given section of plains terrain is characterized predominantly by features typical of one particular agent.

**Features associated with stream erosion.** Valleys and the divides between them characterize plains sculptured largely by stream erosion. Surfaces are usually irregular rather than flat, but differ among themselves in size, form, arrangement, and spacing of their valleys. Since stream-eroded plains are far more widespread than any other class, these differences are especially significant and frequently encountered.

The depth, width, and cross-section form of stream valleys depend upon the relative rates of erosional deepening of the valley by the stream itself and of erosional widening of the valley by the cutting of tributary ravines and by rain wash and slow downward creep of soil on the valley sides, locally aided by undercutting of the slope base by the stream. Narrow, steep-sided valleys are favored by steep, swift-flowing streams that erode rapidly and by an absence of excessive surface runoff down the valley sides from the adjacent uplands. Conversely, broad, gentle-sided valleys are typically associated with slow-flowing streams of gentle gradient; valley walls are composed of weak materials and are heavily washed by runoff. Since streams flowing on plains are usually gentle in gradient, valleys on plains are commonly wide, shallow, and gentle-sloped, but significant exceptions are numerous.

The spacing of valleys on a plain indicates the stage to which development of the stream system has advanced. Normally through the course of time a major stream develops an increasing number and length of tributaries, until eventually the entire drainage basin is occupied by valleys and their side slopes. When this point is reached, tributary growth is complete. Continued development involves the reduction of slopes and lowering of divides between streams until the whole surface has been brought low.

Plains differ greatly in the stage of development their stream systems have reached. Plains on which tributary growth is highly incomplete, so that broad uncut uplands remain between widely spaced valleys, are called youthful. Plains on which tributary growth is complete are termed mature, and are typically rolling surfaces like the Appalachian Piedmont between Washington, D.C.,

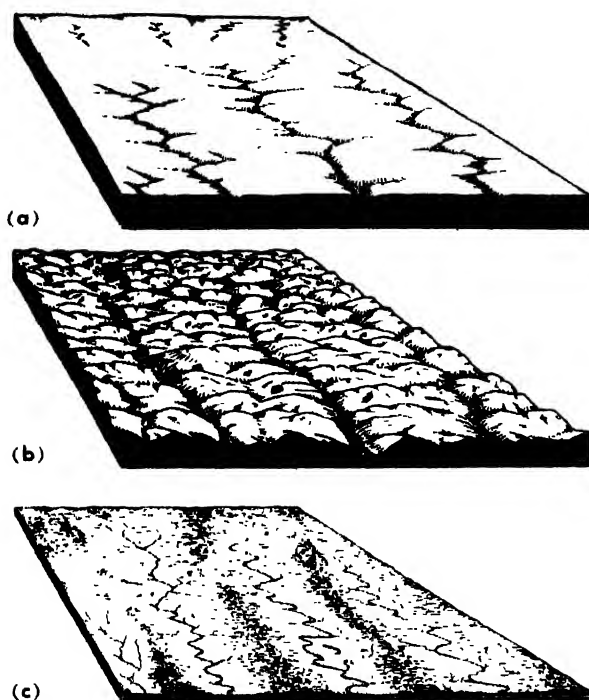


Fig. 1. Ideal stages in the sequence of development of a land surface under the effects of stream erosion. (a) Youth; (b) maturity; (c) old age. (From V. C. Finch, G. T. Trewartha, A. H. Robinson, and E. H. Hammond, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)

and Atlanta, Georgia. Old-age plains (also called peneplains) should, ideally, be low lying and smooth, since they would represent the ultimate of erosion. However, they require so long to develop that they are usually re-uplifted and re-eroded before completion, and hence are rarely found. See FLUVIAL EROSION CYCLE; STREAM TRANSPORT AND DEPOSITION.

The pattern of valleys on plains depends chiefly upon the pattern of outcrop of rock materials of contrasting resistance. In the absence of strong contrasts the pattern is usually branching and tree-like as in Fig. 1. Where there is great differential resistance to erosion, the unusually resistant rocks form drainage divides, whereas weak rock belts are soon excavated into broad valleys or lowlands. Where erosional plains bevel across gently warped

rock strata of varying resistance, the belts of outcrop of the more resistant strata form strips of higher, rougher country, with an abrupt escarpment on one margin and a more gradual dip-slope in the direction toward which the strata are inclined. These *cuestas* are common features in the American Middle West and Gulf Coastal Plain and in western Europe. The various wolds and downs of England and the *côtes* of northeastern France are *cuesta* ridges.

Most plains that develop in dry climates are characterized predominantly by stream-produced landforms, in spite of infrequent rain. The development of valley systems and erosional features follows the same general rules as it does in humid regions. However, some differences in relative rates and relative significance of certain of the developmental processes produce distinctive landscape characteristics in arid lands. First, rock decomposition is very slow, so that the surface accumulation of weathered material is normally thin and coarse textured. Second, the sparse vegetation affords to the naked surface little protection against the battering and washing of the occasional torrential rains. As a result, the upper slopes become strongly gullied and often stripped of much of their covering material, leaving bedrock exposed over wide areas. The debris that is eroded from these upper slopes is rarely carried far, however, because of the short duration and local nature of the rains and hence the intermittent character of stream flow. Therefore most of the debris load is dropped in the neighboring basins and valley floors, "drowning" broad areas beneath plains of silt, sand, or gravel. Hence denuded and gullied upper slopes and broad depositional flats in the lowlands are characteristic features of desert plains. See DESERT EROSION FEATURES.

**Solution-marked plains.** Features resulting from underground solution characterize several rather extensive areas of plains. The principal features of this class are depressions, or sink holes, produced by collapse of caverns underneath. Significant groundwater solution is largely confined to areas underlain by thick limestones. As subsurface cavities are progressively enlarged by solution more and more drainage is diverted to subterranean channels. Surface streams become fewer.

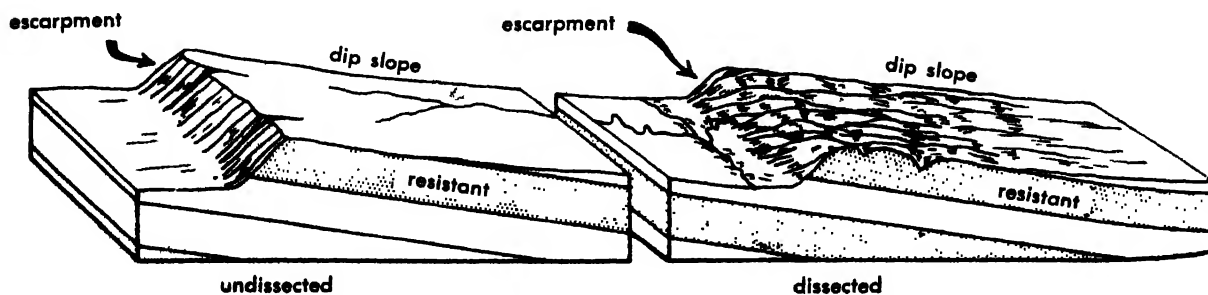


Fig. 2. The structure and surface form of *cuestas*. Dissected form shown at right is especially common. (From V. C. Finch, G. T. Trewartha, A. H. Robinson,

and E. H. Hammond, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)

often disappearing into the ground after a short surface run. Eventually solution cavities near the surface collapse, forming surficial depressions of various sizes. Some are shallow and inconspicuous; others are great steep-walled pits or elongated enclosed valleys. Some of the small depressions contain lakes, because their outlets are plugged with clay. The most extensive areas of solution-featured plains in the United States are in central and northern Florida and in the Panhandle of Texas. In both of these areas shallow sink holes, some lake filled, are numerous. Some of the areas of most active solution work have developed surfaces far too rough to fall under the heading of plains. This is especially true for the mountainous area of great sinks and solution valleys in the Dalmatian Karst of Yugoslavia. See KARST TOPOGRAPHY.

**Patterns associated with stream deposition.** Alluvial features are so called from the term alluvium, which refers to any stream-deposited material. As a group, alluvial plains are among the smoothest and flattest land surfaces known. True stream-deposited plains fall into three classes: (1) floodplains, laid down along the floors of valleys; (2) deltas, formed by deposition at the stream mouths; and (3) alluvial fans, deposited at the foot of mountains or hills.

**Floodplains.** The flat bottomlands so common to valley floors develop where the gentle lower reaches of a stream system have more sediment load fed to them by their tributaries than they are able to carry. This sediment is deposited in the stream bed or, in flood time, across the whole width of the valley floor. Because of continued deposition and choking of the stream bed, repeated flooding, and the ease with which the alluvium can

be moved, streams continually shift their channels on floodplains. On plains of fine silt, the channel is usually highly sinuous or meandering, while on a sandy floodplain the channel is braided, that is, broad, shallow, and intricately subdivided by innumerable sand bars. In either case, many loops or strands of abandoned channels, now mostly dry and filled in, scar the floodplain surface. On silty floodplains the highest land is commonly found immediately adjacent to the channel, while farther back from the stream the surface is slightly lower. These higher strips, called natural levees, are formed by active deposition during floods, when the velocity of flow is abruptly checked and the bulk of the sediment dropped immediately as the water leaves the swift current of the deep channel to spread thinly over the plain.

On major floodplains the ground water table is everywhere close to the surface, and swampy land is common in the abandoned channels and shallow swales behind the natural levees. For this reason the natural levees are especially sought after for cultivation, town sites, and transportation routes. They normally provide the best-drained land on the floodplain and the last to be flooded as the waters rise. Though harassed by a high water table and recurrent floods, floodplains, especially silty ones, are often prized agricultural land, because of the level, easily tilled surface. In some cases the alluvium is also more fertile than the soils of the surrounding uplands.

Here and there along the sides of valley bottoms and somewhat above the present level of stream or floodplain, strips of smooth land extend in the form of benches or terraces made up of alluvium. These are remnants of earlier floodplains that have been largely destroyed by a renewal of down-cutting by their streams. In many instances the stream has then recommenced deposition and has built a new floodplain at a lower level. These alluvial terraces are often valuable agricultural lands and also serve well for town sites and transportation routes because they stand above flood levels.

The surface features of deltas are essentially the same as those just described. Indeed, deltas are often simply the seaward extensions of floodplains. As a rule, however, the delta surface is even less well drained than the floodplain surface and at its outer margin may merge with the sea through a broad belt of marshy land. As the delta grows, continual clogging of the stream mouth produces repeated diversion and bifurcation of the channel. Hence in a well-developed delta, discharge is rarely through a single channel, but through a spreading network of diverging channels, or distributaries.

**Deltas.** These vary greatly in size and form. Many, like those of the Mississippi, the Nile, the Danube, or the Volga, are immense fan-shaped features that have produced broad coastal bulges by their growth. Others, like those of the Colorado, the Po, or the Tigris-Euphrates, though no less extensive, are less apparent on the map because they have been built in large coastal embayments.

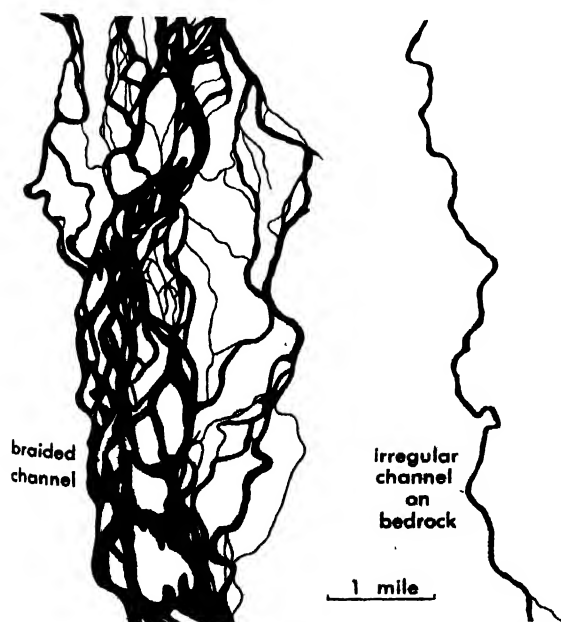


Fig. 3. Characteristic patterns of stream channels on alluvial and on bedrock surfaces. (From V. C. Finch, G. T. Trewartha, A. H. Robinson, and E. H. Hammond, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)

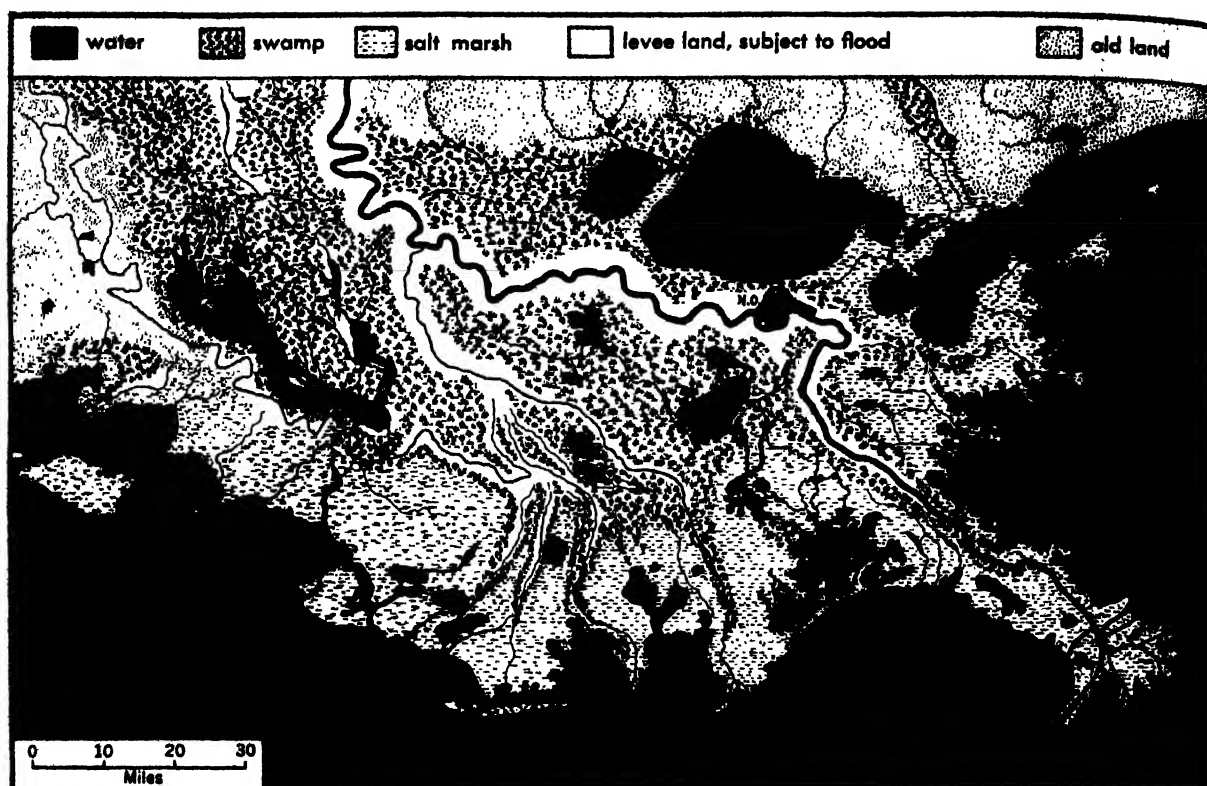


Fig. 4. Swamp, marsh, and natural levee lands in the Mississippi River delta. (From V. C. Finch, G. T. Tre-

wartha, A. H. Robinson, and E. H. Hammond, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)

Some great rivers have no true deltas because their sediment load has been dropped in some interior settling basin, as the Great Lakes remove most of the sediment from the St. Lawrence system and the Congo deposits most of its load in its broad upland basin.

Like floodplains, deltas are sometimes highly valued as agricultural lands, though they have the same problems of poor drainage and frequent flooding. The Nile delta and the huge, silty delta plain of the Hwang Ho, in north China, are famous centers of cultivation. Many deltas, of which that of the Mississippi is a good example, are too swampy to permit tillage except along the natural levees. The Netherlands, occupying the combined deltas of the Rhine and Maas, stands as an example of what can be done toward reclamation of such lands when the need is great.

If a stream emerges upon a gentle plain from a steeply plunging mountain canyon, its velocity is abruptly checked, and it deposits most of its load at the mouth of the canyon. Because of the tendency toward repeated choking and diversion of the channel, the deposit assumes the form of a broad, spreading fan, essentially similar to a delta, even to the diverging distributary channels. Usually, however, the gradients developed are steeper than those on a delta, especially near the head of the fan, where the coarsest sediment is to be found.

**Alluvial fans.** Small individual alluvial fans are common features in mountainous country, especially where the climate is dry except for occa-

sional torrential showers. Particularly significant, however, are the occurrences, along the foot of long, precipitous mountain fronts, of rows of alluvial fans that have coalesced to form an extensive, gently sloping piedmont alluvial plain. Such surfaces sometimes achieve great areal extent. The city of Los Angeles is built on such a plain. Still larger ones occupy much of the southern part of the Central Valley of California and stretch eastward from the Andes in northwestern Argentina, Paraguay and eastern Bolivia.

Because of their smoothness and ease of tillage, alluvial fans, like other alluvial surfaces, are often especially amenable to cultivation. They are particularly significant in drier areas, partly because of the ease with which water may be conducted by gravity from the mountain canyon to any part of the fan, and partly because the thick, porous alluvium itself serves as a reservoir in which ground water is naturally stored.

Closely allied to stream-deposited plains are lacustrine, or lake-bottom plains, and newly emerged coastal plains, which are recently exposed areas of the former shallow sea bottom. These are the nearly featureless surfaces of sedimentary deposits that have been carried into the body of water by streams or by wave erosion and further smoothed by action of waves and currents. Former beach lines and other shore features, sometimes in multiple sets, often form the only noteworthy breaks in the monotonous flatness. In some places, shallow valleys have been cut by streams since the surface

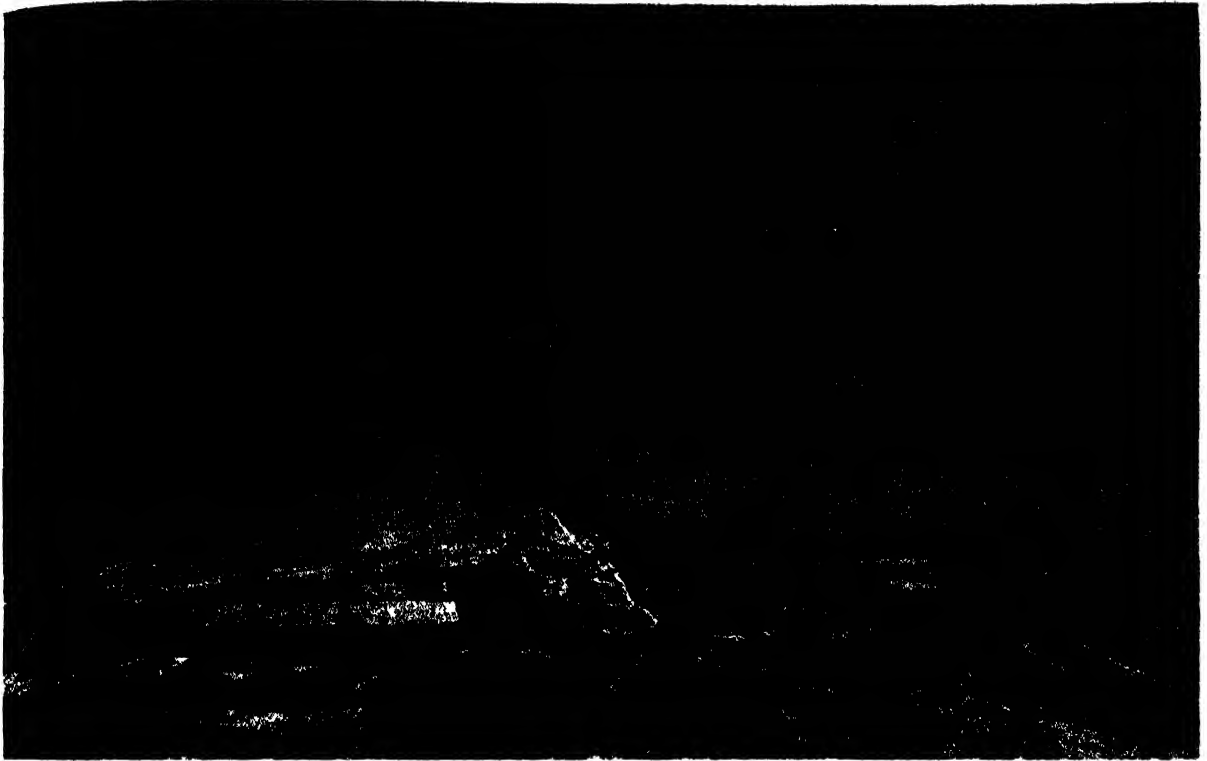


Fig 5 Alluvial fans coalescing to form a piedmont alluvial plain near Independence, Calif. (Spence Air Photos)

became exposed. The lower parts of former lake floors and the outer margins of coastal plains are often conspicuously poorly drained.

The flat surfaces upon which Detroit, Toledo, Chicago, and Winnipeg stand are all lacustrine plains (Fig. 7), as are the famed Bonneville Salt Flats of western Utah. The south Atlantic and Gulf margins of the United States and much of the Arctic fringe of Siberia are examples of newly emerged coastal plains. Some of these plains represent valuable agricultural land; others are excessively swampy or sandy. See DELTA; FLOOD PLAINS.

**Character from recent glaciation.** Plains recently glaciated assume distinctive associations of

forms, in accordance with the corresponding variety of erosional and depositional processes involved. As a group, plains owing their surface features largely to the work of continental ice sheets are distinguished by the absence of a systematic integrated pattern of streams, valleys, and divides; by the presence of great numbers of lakes and swamps; and by the occurrence of surface materials obviously not derived from the local bedrock.

Although there were probably four major periods of glacial growth and decay during the Ice Ages, only those areas covered during the last glacial stage (the Wisconsin glaciation) show distinctively ice-produced surface forms. The forms produced

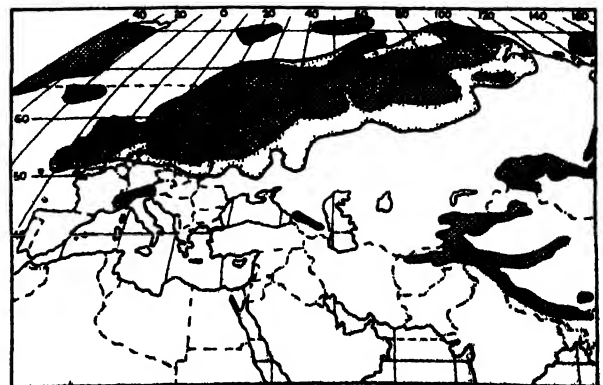
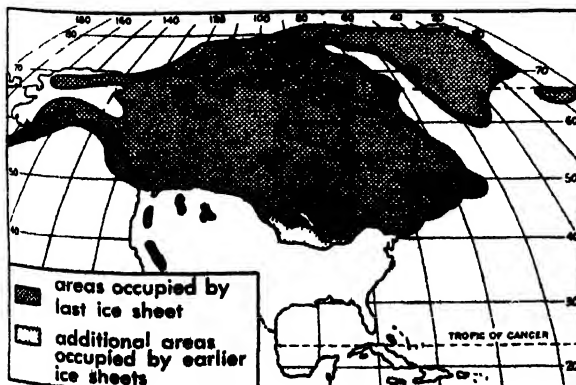


Fig 6 Map showing areas covered by former continental glaciers. (After R. F. Flint from V. C. Finch,

G. T. Trewartha, A. H. Robinson, and E. H. Hammond, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)



elsewhere by earlier glaciations have been largely obliterated by stream erosion and soil creep. The last glaciation, at its maximum, covered all of Canada, the Great Lakes states, and New England as well as Scandinavia, Finland, most of the British Isles, Baltic Europe, northern Russia, and northwestern Siberia. These are the areas of glacial plains.

Northern and eastern Canada and the Fennoscandian upland, the areas in which the ice sheets developed and from which they spread out, now exhibit more evidence of glacial erosion than of glacial deposition. Over broad areas the soil has been largely stripped off, exposing patches and knobs of scoured bedrock. Shallow depressions have been eroded in the underlying rock, and patchy deposits of glacially transported debris (drift) are strewn thinly about over the surface. Lakes and swamps of irregular shape are extremely abundant. Some occupy the eroded hollows, others accumulating where streams have been blocked by drift deposits. Streams wander from lake to lake, with many waterfalls and rapids along their devious courses.

These areas of dominant glacial erosion owe their existence to two factors. First, the unusually resistant rock that happens to underlie both areas did not yield large quantities of drift. Second, most of the drift that was eroded from these regions was carried out toward the outer edges of the ice sheets, where melting permitted it to be dropped. See EROSION; GLACIATED TERRANE.

Because of their patchy, thin, and stony soils and large areas of standing water, the glacially scoured regions would not be favorable areas for human occupancy even if the climate were less severe than it is. Even the coniferous forests native to the areas are neither dense nor luxuriant.

The outer parts of the glaciated areas, on the other hand, are characterized chiefly by features of glacial deposition, though erosional features are not rare. Throughout these areas in western Canada, the north-central United States, and northern Europe outside of the Fennoscandian upland, glacial drift was strewn over the preexisting terrain in a sheet of irregular thickness and varied composition.

*Till.* Much of the drift represents mixed rock and soil material deposited beneath the ice or at the edge of the sheet directly by melting ice. This material, called till, is as a rule most thickly deposited in the valleys, and thinly over the ridge tops, thus having the effect of reducing terrain irregularity. The surface of the till sheet itself is usually gently rolling, with many shallow depressions containing lakes or swamps, numerous haphazardly placed swells and hillocks, and no systematically arranged stream valleys. Hummocky, often stony ridges, called marginal moraines, mark places where the fluctuating edge of the ice remained stationary for long periods of time. In several localities, notably in eastern Wisconsin and western New York, are swarms of smooth, low drift hills,

all elongated in the direction of ice movement. The mode of origin of these drumlins is uncertain.

The surfaces of stony till plains are usually more irregular than those on clay till. Northeastern Illinois has a remarkably smooth surface developed on clay till, apparently eroded from the Lake Michigan basin. Eastern Wisconsin, northern Michigan, western New York, and southern New England, on the other hand, have more rolling surfaces underlain by till having a high content of stone and sand. In a few areas, especially in southern New England and in the marginal moraines elsewhere, the till is so very stony as to impede cultivation.

*Outwash.* Some of the debris transported by the ice is carried out beyond the glacial margin by streams of meltwater. This material, called outwash, may be deposited as a floodplain (here called a valley train) along a preexisting valley bottom, or it may be spread broadcast over a preexisting plain in a form similar to an alluvial fan. In either case the surface will usually be smooth, with the features that are typical of such alluvial plains. Unlike the heterogeneous, unsorted and unstratified till, outwash material is usually distinctly layered and well sorted in size. The fine material of silt and clay size is carried out downstream, leaving the coarser sands and gravels to form the outwash deposits. Most of the gravel and sand pits that abound in glaciated areas are developed in outwash plains.

Whereas much outwash was deposited beyond the extreme limits reached by the ice, some was also laid down over already deposited till surfaces after the ice had melted back from its maximum extent. Under such conditions the surface of the outwash plain is sometimes pitted and lake strewn, the depressions having formed as a result of the melting of relict ice masses that were buried by outwash deposition.

Patches and ribbons of outwash are common in glaciated areas, and in a few places, into which unusual quantities of meltwater were funneled, there are very extensive sandy plains. Noteworthy in this respect are the southern Michigan and northern Indiana area and Europe immediately south of the Baltic.

*Lacustrine plains.* Also present in and around the glaciated areas are numerous lacustrine plains marking the beds of former lakes that resulted from the blocking of rivers by the glacial ice itself. During the melting of the last ice sheet, while the St. Lawrence and other northward-flowing rivers were still ice dammed, the Great Lakes basins were much fuller than now, and overflowed to the southward. When lake levels were eventually lowered, lacustrine plains were exposed, notably about Chicago, at the western end of Lake Erie, and about Saginaw Bay in Michigan. One of the most featureless plains of North America occupies the former bed of an immense lake (Lake Agassiz) that was present in late glacial time in southern Manitoba, northwestern Minnesota, and eastern North Dakota.



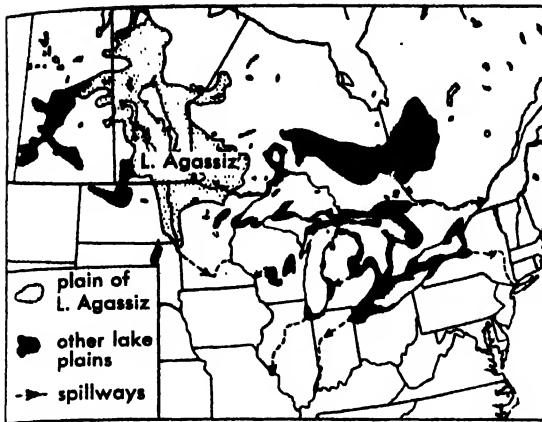


Fig. 7. The plain of former Lake Agassiz and other glacial lake plains in North America. (From V. C. Finch, G. T. Trewartha, A. H. Robinson, and E. H. Hammond, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)

**Patterns reflecting wind action.** Features formed by the wind are less widespread and, as a rule, less obtrusive than the forms produced by streams and glaciers. Since the wind can attack only where the surface is almost free of vegetation, its work is strongly evident only in arid regions and in the vicinity of beaches and occasionally exposed river beds in more humid lands. Today plowed fields are also important prey for wind erosion.

The most striking and significant wind-produced features are sand dunes. Where sand is exposed to strong winds, it is moved about for short distances and accumulates in heaps in the general vicinity of its place of origin. Sand dunes assume many forms, from irregular mounds and elongated ridges to crescent-shaped hills and various arrangements of sand waves, depending apparently upon the supply of sand, the nature of the underlying surface, and the strength and directional persistence of the wind.

Nearly all the truly extensive areas of sand dunes are found in the Eastern Hemisphere, especially in the Sahara, Arabia, central Asia, and the interior of Australia. Most of the dunes have been whipped up from alluvium that has been deposited in desert basins and lowlands.

In several regions of the world, notably in north-central Nebraska and in the central and western Sudan south of the Sahara, extensive areas of dunes have become covered by vegetation and fixed in position since they were formed, suggesting the possibility of climatic change.

Other wind-formed features are polished and etched outcrops of bedrock that show the effects of natural sand-blasting, gravel "pavements" resulting from the winnowing out of finer material from mixed alluvium, and shallow blowouts, which are depressions formed by local wind erosion.

The finer silty material moved by the wind is spread as a mantle over broad expanses of country downwind from the place of origin. Though usually

thin, this mantle in places reaches a thickness of a few tens of feet, thus somewhat modifying the form of the surface. The extensive deposits of unstratified, buff-colored, lime-rich silty material known as loess are believed to have originated from such wind-laid deposits. Loess is abundant in the central United States, eastern Europe and southern Russia, and in interior north China. It yields readily to gully erosion and has the facility of maintaining remarkably steep slopes, so that erosional terrain developed on deep loess is often unusually rough and angular. See LOESS.

#### INTERRUPTED PLAINS

Interrupted plains, or plains broken by some features of considerable relief, occur widely and merit independent treatment. They may be divided into two contrasting groups: (1) tablelands, which are upland plains deeply cut at intervals by steep-sided valleys or broken by escarpments, and (2) plains with spaced hills or mountains, in which the excessive relief is afforded by steep-sided eminences that rise above the plain. Both types of surface, with their combination of plain and rough land, suggest histories of development that combine extensive gradation and strong tectonic activity.

**Tablelands.** These are essentially youthful plains that have been unusually deeply cut by valleys. This requires that the plain shall have been brought to a level hundreds or even a few thousands of feet above the level to which streams can erode. In most cases, this elevation has been accomplished by the broad uplifting of an erosional or alluvial plain, but in a few instances the plain has been built to high level by the deposition of many thick sheets of lava.

It is also necessary that, during the time required for a few major streams to have cut deep canyons, large areas of the upland plain shall have suffered no significant dissection from tributary development, a condition requiring special circumstances. Such inhibited tributary growth is usually the result of either (1) slight local surface runoff of water because of aridity, extreme flatness of the upland, or highly porous material; or (2) the pres-

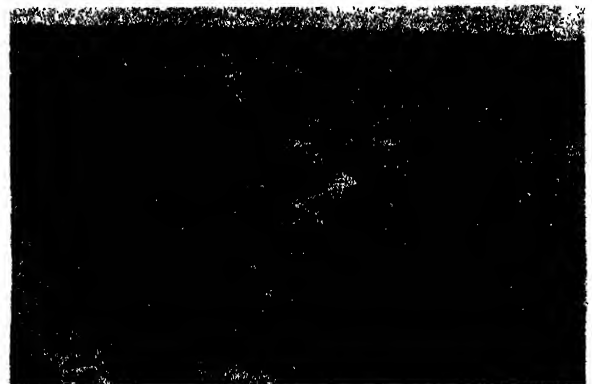


Fig. 8. An ideal tableland. Canyon de Chelly National Monument, northeastern Arizona. (Spence Air Photos)

ence, at the upland level, of very resistant strata of rock that have permitted only the most powerful streams to cut through. Hence tablelands are predominantly a dry-land terrain type. Those that do occur in rainier climates usually indicate the presence of an exceptionally resistant "cap-rock" layer.

The cliffs and escarpments that are common features of tableland regions are sometimes fault scarps, produced by the breaking and vertical displacement of the crust during uplift. More often, however, they are simply steep valley sides that have retreated long distances from their original positions under the attack of weathering and erosion. Sometimes a once extensive tableland will have been so encroached upon by the wearing back of its bordering escarpments that nothing remains but a small, flat-topped mesa or butte.

Tablelands are the least widespread of the major terrain types, presumably because of the restrictive circumstances under which they can develop. The most extensive examples occur in the American continents. Especially noteworthy are the Colorado Plateaus, largely in northern Arizona and southern Utah, which represent a complex erosional plain that has been greatly uplifted and then deeply carved by the Colorado River and its major tributaries. Preservation of large sections of the upland plain has been favored by dryness and by the presence of nearly horizontal resistant rock strata. The few major streams that have cut deep canyons are all fed from moister mountainous areas round about.

The Columbia Plateau in eastern Washington is a porous lava plain, cut by the Columbia River and a few tributaries. The northern Great Plains, lying east of the Rocky Mountains from Nebraska northward into Alberta, are an old alluvial plain, now crossed by valleys of moderate depth that have been cut by streams issuing from the Rockies. The Patagonian Plateau of southern South America is a somewhat similar surface, locally reinforced by extensive lava flows. Parts of the upland of interior Brazil, though in a moist environment, retain a tableland form because of the resistance of thick sandstone beds at the upland level.

Though there are significant exceptions, tablelands as a group suffer, in their economic development, from the difficulty of passage through their narrow gorges and across their numerous escarpments, and in most cases also from the dryness of their upland surfaces.

**Plains with spaced hills or mountains.** These are much more widespread than tablelands and largely represented in each continent. Surfaces included under the general heading vary from plains studded with scattered small hills and hill groups to mountain-and-plain country in which high, rugged ranges occupy almost as much space as the plains between them.

Surfaces of the general type can be produced by two quite different lines of development. (1) They may be high-relief lands that have been brought to the erosional stage of early old age, in which case

the isolated hills represent the only remnants of a once extensive highland. (2) They may represent areas in which separated hills or mountains have been constructed by volcanic eruption or by folding or buckling of the crust, the land between them having remained smooth from the outset, or having been smoothed by erosion and deposition since the mountains were formed.

Examples of the first course of development are limited in the United States to small areas in southern New England and in the Appalachian Piedmont just east of the Blue Ridge. Many patches occur in the glacially scoured sections of northern and eastern Canada and similarly in northern Sweden and Finland. Extensive areas are found in the southern regions of Venezuela and Guiana, in the upland of eastern Brazil, and especially on the uplands of central and southern Africa.

In such areas the plains are typical late-stage erosional-depositional surfaces, usually gently undulating. The remnant hills (monadnocks) rise abruptly from the plain, like islands from the sea. Although monadnocks commonly represent outcrops of unusually resistant rocks that have withstood erosion most effectively, many are not thus distinguished, but owe their existence solely to their position at the headwaters of the major stream systems, where they are the last portions of the highland to be reduced.

Surfaces on which spaced mountains have been constructed by crustal deformation or volcanic activity occur extensively in the great cordilleran belts of the continents, and rarely outside of those belts. The largest of all such regions is the Basin-and-Range section of western North America, which extends without interruption from southeastern Oregon through the southwestern United States and northern Mexico to Mexico City. The majority of the rugged mountain ranges of this section are believed to represent blocks of the crust that have been uplifted or uptilted and then strongly eroded. The plains between them are combination erosional-depositional surfaces, some of which have clearly expanded at the expense of the adjacent mountains. As the mountains have been reduced by erosion, there have evolved at their bases smooth, gently sloping plains that are in part erosional pediments, closely akin to old-age peneplains, and in part piedmont alluvial plains. Many of the basins have interior drainage, and in these the floor is likely to be especially thickly alluviated and to contain a shallow saline lake or alkali deposit in its lowest part.

Other areas of somewhat similar terrain are found in the central Andes, in Turkey and the Middle East, and in Tibet and central Asia. The Tibetan and central Andean sections are noteworthy for the extreme elevation (12,000–15,000 ft) of their basin floors. In general the ranges and basins in the Asiatic areas are developed on a grander scale than those in North America.

It is a curious circumstance that practically all these regions are dry and exhibit the intensely eroded mountain slopes and alluvially drowned

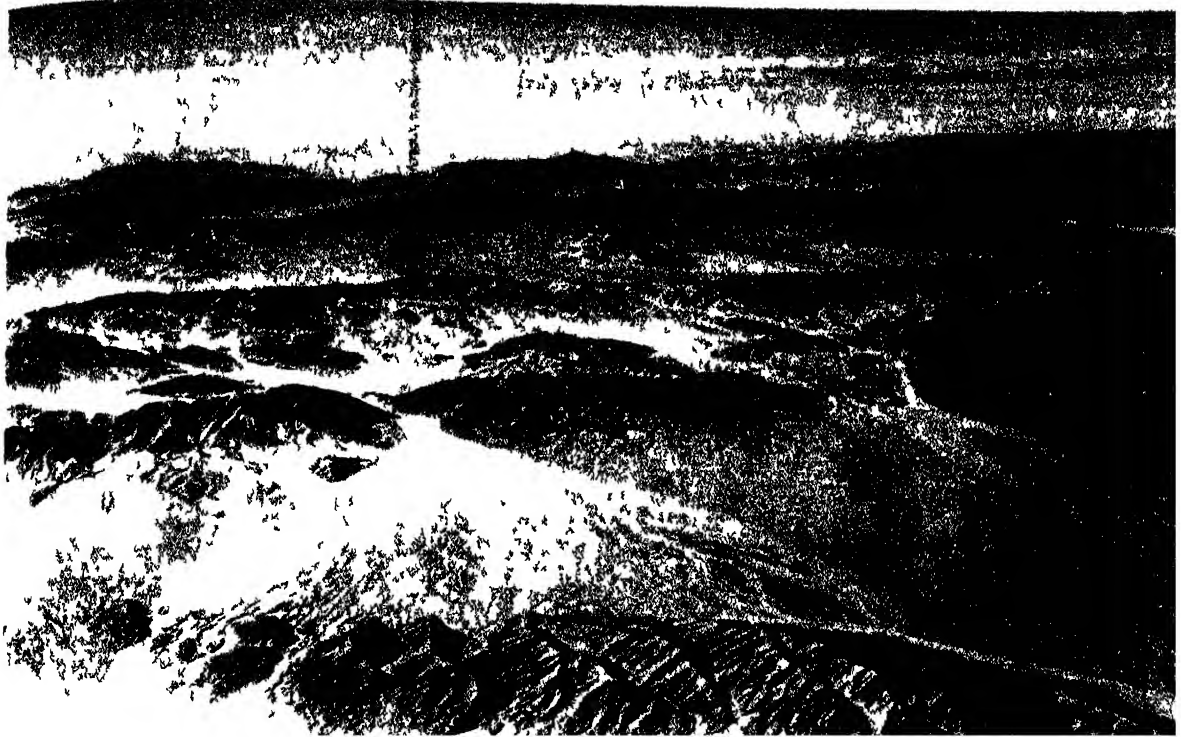


Fig 9 Basin and range country in the Mojave Desert, southeastern California (Spence Air Photos)

Basin floors characteristic of that climatic realm. For this reason they are only locally useful to man, in spite of the large amount of smooth land that they afford. There are, however, many important oases usually near the bases of the mountains or along the courses of the few streams that have wandered in from moister adjacent regions.

[F. H. HAMMOND]

*Bibliography.* V. C. Finch, G. T. Trewartha, A. H. Robinson, and F. H. Hammond, *Elements of Geography*, 4th ed., 1957.

## Planck's constant

A fundamental physical constant which represents the elementary quantum of action, action being defined as energy multiplied by time. Introduced by Max Planck in 1900, it has the value  $h = 6.6256 \times 10^{-27}$  erg-sec or  $6.6256 \times 10^{-34}$  joule-sec. The symbol  $\hbar$ , sometimes called the Dirac  $\hbar$ , is often used for convenience in physics to denote the quantity  $h/2\pi$ , where  $\pi = 3.1416$ .

The unique feature of Planck's constant is this: as used by Planck in deriving his radiation law,  $h$  multiplied by the frequency of radiation represented a bundle of energy, that is, a quantum of energy. Radiant energy at any wavelength can occur only as multiples of this energy; thus energy is quantized. This was a fundamental departure from the beliefs of physics up to Planck's time and was indeed quite startling to Planck himself. Until Planck deduced his law to satisfy experimental data the general belief was that energy could be divided indefinitely. The quantization of energy implied by Planck's constant laid the foundations of quantum physics upon which much of modern physics has

flourished. The frequency  $\nu$  of emitted radiation is related to the quantized energy,  $\Delta E$ , by the relation  $\Delta E = h\nu$ .

The concept of energy quantization first began to win general acceptance after 1905 when Albert Einstein showed that it gave a good account of some of the then puzzling features of the photoelectric effect. Later A. H. Compton showed that the electromagnetic quanta also carry momentum  $p$  which is related to the wavelength  $\lambda$  by  $p = h/\lambda$ .

For an extended discussion of Planck's radiation law see HEAT RADIATION. For a discussion of the role of Planck's constant in theoretical physics, see QUANTUM MECHANICS. See also ATOMIC CONSTANTS.

[H. G. SELL; P. J. WALSH]

## Planck's radiation law

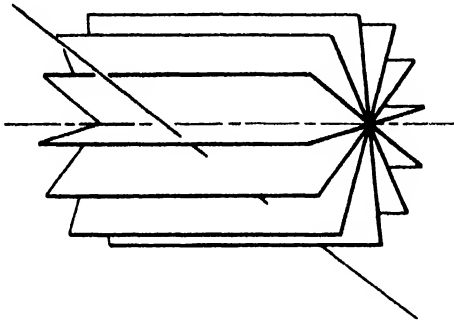
A law of physics which gives the spectral energy distribution of the heat radiation emitted from a so-called black body at any temperature. Discovered by Max Planck early in the twentieth century, this law laid the foundation for the advent of the quantum theory, because it was the first physical law to postulate that electromagnetic energy exists in discrete bundles, or quanta. For an extended discussion, see HEAT RADIATION. See also QUANTUM MECHANICS.

[H. G. SELL; P. J. WALSH]

## Plane

In the euclidean definition, a surface is "that which has length and breadth only," and a plane surface is "a surface which lies evenly with the straight lines on itself." A definition of a plane that avoids this somewhat loosely defined concept of a surface

is as follows: "A plane is a set of at least three points, not all collinear, such that if any two points *A*, *B* of the set are given, all points of the line *AB* are in the set, and such that if any four points *A*, *B*, *C*, *D* of the set are given, then at least one of the pairs of lines *AB* and *CD*, or *AC* and *BD*, or *AD* and *BC* have a common point." (This last restriction prevents a plane from occupying all of space.) Still another definition is this: "A plane is the locus of points each equidistant from two points in space." Points or lines in the same plane are called coplanar. Each two planes are congruent. Two planes that have a common point have a common line. (This is not true in spaces of more than three dimensions.) There is one and only one plane that contains (a) three given noncollinear points, or (b) a given line and a given point not on the line, or (c) two given intersecting lines, or (d) two given parallel lines.



Coaxial planes and an intersecting line.

Although a plane is unlimited in extent, it is usually represented by a parallelogram or other plane figure in a drawing. Three or more planes that have a line in common are called coaxial planes (see illustration). See ANALYTIC GEOMETRY; GEOMETRY, EUCLIDEAN; LINE; POINT. [J.S.F.]

## Plane table

A tripod-mounted drawing board on which topographic survey details are compiled in the field. The table and affixed manuscript sheet are oriented at a point represented by a point on the sheet. Stadia sightings with an alidade locate additional points by direction-distance observations. Points also are located by intersections of the sightings from two different plane-table points. Sufficient additional points are observed to permit sketching planimetric details. Elevation differences also are observed for contour sketching. See ALIDADE; TOPOGRAPHIC SURVEYING AND MAPPING. [R.H.DO.]

## Planer

A machine for the shaping of long, flat, or flat contoured surfaces by reciprocating the workpiece under a stationary single-point tool or tools. Usually the workpiece is too large to be handled on a shaper.

Planers are built in two general types, open-side or double housing. The former is constructed with

one upright or housing to support the crossrail and tools. The double-housing type has an upright on either side of the reciprocating table connected by an arch at the top.

Saddles on the crossrail carry the tools which feed across the work. A hinged clapper box, free to tilt, provides tool relief on the return stroke of the table. A variation is the milling planer; it uses a rotary cutter rather than single-point tools. See SHAPER; see also MACHINING OPERATIONS; WOODWORKING. [A.T.]

## Planet

A small, solid celestial body circulating around a star, in particular our Sun. Besides Earth, the eight known main planets of the solar system are Mercury, Venus, Mars, Jupiter, Saturn, Uranus, Neptune, and Pluto; in addition, over 1600 minor planets, or asteroids, circulating mainly between the orbits of Mars and Jupiter, are known.

**Classification.** There are two main groups of planets: the small terrestrial planets—Mercury, Venus, Earth, Mars, and Pluto—and the large or major planets—Jupiter, Saturn, Uranus, and Neptune. The asteroids may be the remnants of a very small planet (or planets) of the terrestrial group. Each of the main planets from the Earth\* to Neptune is accompanied by one or more secondary planets or satellites. Pluto may once have been a satellite of Neptune.

The planets are also divided into interior planets, Mercury and Venus, circulating inside the Earth's orbit, and exterior planets, from Mars to Pluto, circulating outside it.

**Kepler's laws.** The motions of the planets in their orbits around the Sun are governed by three laws discovered by J. Kepler at the beginning of the seventeenth century.

First law: The orbits of the planets are ellipses of which the Sun occupies a focus.

Second law (law of areas): Equal areas of the ellipse are described by the radius vector from the Sun to the planet in equal intervals of time.

Third law (harmonic law): The squares of the periods of revolution *P* are proportional to the cubes of the major axes of the orbits *2a*; that is, for all planets the ratio  $P^2/a^3$  is equal to a constant. The ratio is equal to unity if *a* is in astronomical units and *P* in sidereal years. One astronomical unit (AU) is the mean distance from Earth to the Sun and is approximately equal to  $93.5 \times 10^6$  mi.

The constant of the harmonic law is given by Newton's law of gravitation as  $G(M + m)/4\pi^2$ , where *M* and *m* are the masses of the Sun and the planet, and *G* is the constant of gravitation. See GRAVITATION.

Kepler's laws are true to a first approximation only when the mutual perturbations of the motions of the planets by the others are neglected.

**Bode's law.** From Kepler's third law, the relative mean distances (semimajor axes *a*) of the planets to the Sun can be derived from their ob-

served periods of revolution  $P$ . These relative distances obey approximately an empirical law discovered by J. D. Titius and published by J. E. Bode in 1772; it may be expressed by the formula  $a = 0.4 + 0.3 \times 2^n$ , if  $a$  is in AU, and  $n$  is set equal to  $-\infty$  for Mercury, 0 for Venus, 1 for Earth, 2 for Mars, and so forth. The computed and observed distances are given in Table 1.

Although Bode's law was used in the prediction of the orbit of Neptune, the actual distance falls short of the theoretical value, and the discrepancy is still greater in the case of Pluto (see Fig. 1).

**Planetary configurations.** In the course of their motions around the Sun, Earth and other planets occupy a variety of relative positions or configurations (Fig. 2), the principal of which are designated as follows: the interior planets are in conjunction with the Sun when closest to the Earth-Sun direction, either between the Earth and the Sun (inferior conjunction) or beyond the Sun (superior conjunction). On rare occasions when the planet is very close to the plane of the Earth's orbit at the time of an inferior conjunction, a transit in front of the Sun is observed. See TRANSIT (ASTRONOMY).

Between conjunctions, the angular distance from the planet to the Sun, or the elongation, varies up to a maximum value; the greatest or maximum elongations of Mercury and Venus are  $28^\circ$  and  $47^\circ$ , respectively. The exterior planets are not so limited, and their elongations can reach up to  $180^\circ$  when they are in opposition with the Sun; when the elongation is  $\pm 90^\circ$ , these planets are in quadrature (eastern or western) with the Sun.

The telescopic aspect of the disks of the planets varies according to their configurations, which determine the angle between the directions of illumination and observation, or the phase angle. Between inferior conjunction and greatest elongations, the interior planets show crescent phases, like the Moon between new moon and first or last quarters; between greatest elongations and superior conjunction they show a gibbous phase, like the Moon between quarters and full moon. At superior conjunction, they show a circular disk, fully illuminated and seen face on, while during transits, the dark side is profiled against the Sun. The exterior planets show their full phase at both conjunction and opposition and a gibbous phase near quadrature, when the illumination defect is maximum.

Table 1. Distances of the planets from Sun (AU)

Planet	Bode's law	Observed
Mercury	0.4	0.39
Venus	0.7	0.72
Earth	1.0	1.00
Mars	1.6	1.52
Asteroids	2.8	2.9 (mean)
Jupiter	5.2	5.20
Saturn	10.0	9.55
Uranus	19.6	19.2
Neptune	38.8	30.1
Pluto	77.2	39.5

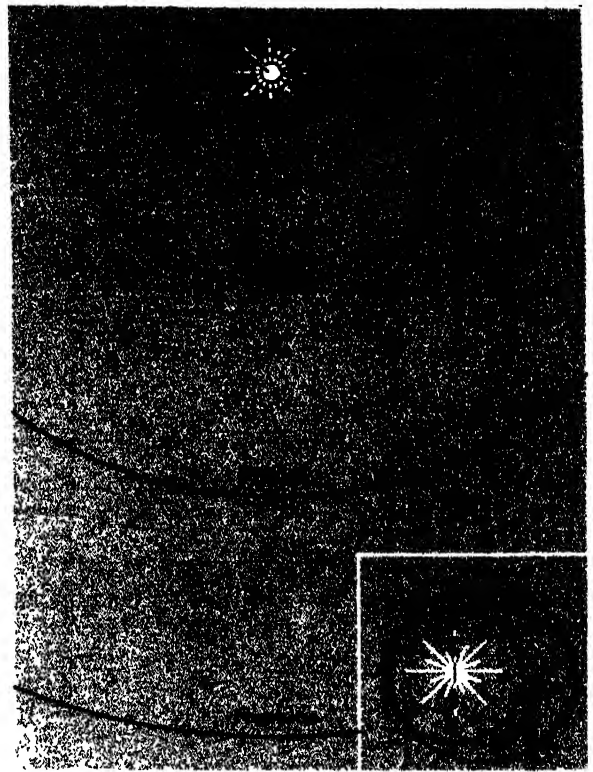


Fig. 1. Plan of the solar system. (From L. Rudaux and G. de Vaucouleurs, *Larousse Encyclopedia of Astronomy*, Prometheus Press, 1959)

**Apparent motions.** The combinations of the orbital motions of Earth and of any other planet give rise to complicated apparent motions of the planets as observed from the Earth. Because the orbits of the main planets are, except for Pluto, little inclined from the plane of the orbit of Earth, the apparent paths of the planets (except Pluto) are restricted to the zodiac, a belt  $16^\circ$  wide centered on the ecliptic. The ecliptic is the path in the sky traced out by the Sun in its apparent annual journey as the Earth revolves around it (see ASTRONOMICAL COORDINATE SYSTEMS). Along this path, the apparent motions of the interior planets with respect to the Sun are alternatively westward, from greatest elongation through inferior conjunction to greatest elongation, then eastward, from greatest elongation through superior conjunction to greatest elongation. The motion of the exterior planets is always westward (see Fig. 3).

The apparent motions with respect to the celestial sphere, that is, to the fixed stars, appear for the interior planets as oscillations back and forth about the position of the Sun steadily moving eastward among the stars. For the exterior planets, the apparent motion is generally eastward or direct, but for short periods near the time of opposition it is westward or retrograde. At times when the direction of the apparent motion on the sphere changes, the planet appears to be stationary.

The mean interval of time between successive returns to the same place with respect to the stars is the sidereal period, which governs the true mo-

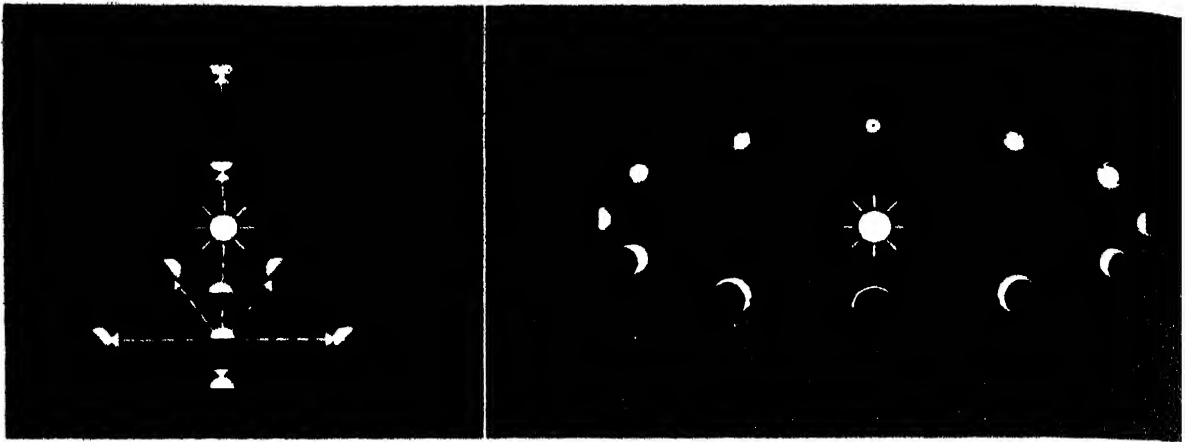


Fig. 2. (a) Planetary configurations. (b) Phases of interior planets. (From L. Rudaux and G. de Vaucou-

leurs, Larousse Encyclopedia of Astronomy, Prometheus Press, 1959)

tion of revolution of the planet on its orbit around the Sun. The mean interval of time between successive returns of the same configuration with respect to the Sun (for example, conjunctions or oppositions) is the synodic period, which governs the apparent motion of the planet as seen from Earth (see Table 2).

**Elliptic motion.** The motion of a planet having an elliptical orbit of semimajor axis  $a$ , with the Sun at the focus  $S$ , brings it every half-revolution to the perihelion  $A$  and to the aphelion  $A'$ , the points of the orbit respectively nearest to and farthest from  $S$ . If  $C$  is the center of the ellipse, the semimajor axis is  $a = CA = CA'$ ; the eccentricity of the ellipse is  $e = CS/CA = CS/a$ , whence  $SA = a(1 - e)$ ,  $SA' = a(1 + e)$  (see ELLIPSE). The distance  $SP = r$  of the planet to the Sun at any other point is  $r = a(1 - e^2)/(1 + e \cos \theta)$ , where the angle  $\theta = \angle ASP$  is the true anomaly. If  $P'$  is the point on the principal circle of radius  $CA = a$  whose projection in the ellipse is  $P$  (see Fig. 4), the eccentric anomaly is the angle  $\theta' = \angle ACP'$ , so that  $r = a(1 - e \cos \theta')$ . If the planet was at perihelion at time  $T$  and returns to it at time  $T + P$ , the mean angular velocity (or mean motion) is  $n = 2\pi/P$ , and the mean anomaly at any time  $t$  is  $M = n(t - T)$ .

The relation between the mean and eccentric anomalies,  $\theta' - e \sin \theta' = M$ , is known as Kepler's equation; its solution gives  $\theta'$  and, consequently,  $r$  at any time  $t$  when the orbital elements  $a, e, n, T$  are known.

**Orbital elements.** The position of a planet in its orbit and the position of the orbit in space is completely defined by seven orbital elements (see Fig. 5): (1) the semimajor axis  $a$ , (2) the eccentricity  $e$ , (3) the inclination  $i$  of the plane of the orbit to the plane of the ecliptic, (4) the longitude  $\Omega$  of the ascending node  $N$ , (5) the angle  $\omega$  from the ascending node  $N$  to the perihelion  $A$ , (6) the sidereal period of revolution  $P$ , or the mean (daily) motion  $n = 2\pi/P$ , and (7) the date of perihelion passage  $T$ , or epoch  $E$ .

If the plane of a planet's orbit is inclined to the plane of the ecliptic, their intersection  $NN'$  is the

line of nodes; in its motion, the planet crosses the plane of the ecliptic from south to north at the ascending node  $N$  and from north to south at the descending node  $N'$ . The longitude of the ascending node is the angle  $\Omega = \angle \Upsilon SN$ , measured in the plane of the ecliptic from the vernal equinox  $\Upsilon$ . The longitude of perihelion is  $\tilde{\omega} = \Omega + \omega = \angle \Upsilon SN + \angle NSA$ , the second angle being measured in the plane of the planet's orbit (see Fig. 5). The location of the plane of the orbit in space is defined by  $i$  and  $\Omega$ , the orientation of the ellipse in this plane by  $\omega$ , its form by  $e$ , and its size by  $a$ , the position of the planet on the ellipse by  $P$  and  $T$  (and by the time  $t$ ). See ORBITAL MOTION.

**Determination of orbital elements.** Accurate observations of the positions of the planets with re-

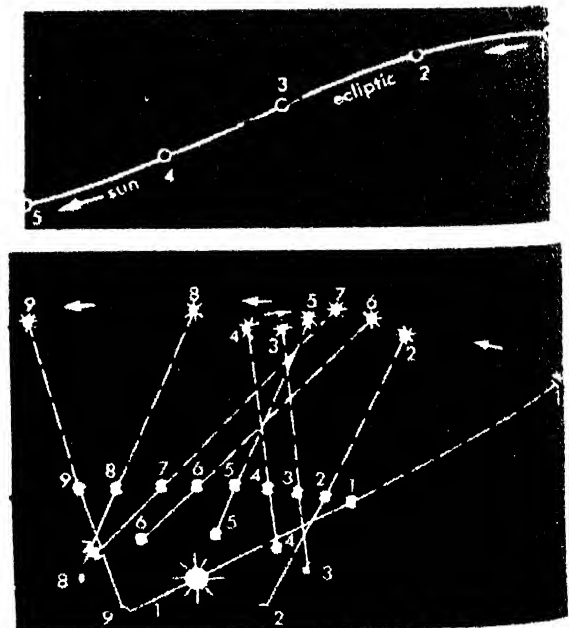


Fig. 3. Apparent motions (a) of an interior planet with respect to the Sun and (b) of an exterior planet with respect to the fixed stars. (From L. Rudaux and G. de Vaucouleurs, Larousse Encyclopedia of Astronomy, Prometheus Press, 1959)



spect to the stars (for example, as measured on photographs) or with respect to the celestial coordinates (for example, by means of the meridian circle instrument) are used to determine the elements of their orbits. In principle, three observations of two coordinates (right ascension and declination) and the laws of elliptic motion are sufficient to determine the six independent elements of a planetary orbit, since by Kepler's third law,  $a^3 \propto P^2$ . In practice, as many observations as possible are combined, and the equations solved by the method of least squares (see LEAST SQUARES, METHOD OF); the elements for a given epoch so obtained are subject to variations and corrections allowing for planetary perturbations. Tables of the motions of the planets for several centuries past and future have been established, from which the yearly ephemerides are extracted in a form convenient for immediate use. See EPHEMERIS; NAUTICAL ALMANAC

The elements of the planetary orbits are given in Table 2

**Planetary sizes.** The apparent diameter of a planet may be determined visually by means of a filar micrometer or preferably, a birefringent or double-image micrometer attached to a telescope, or it may be measured on large-scale photographs. If the apparent diameter of a planet is  $d''$  when its distance to the Earth is  $\Delta$ , the linear diameter is  $D = \Delta \sin d'' = \Delta d''/206,265$ , where  $d''$  is measured in seconds of arc. The linear diameter is expressed in the same units as  $\Delta$ , which is given by the ephemerides in astronomical units; conversion to miles or kilometers is given by the adopted value of the astronomical unit:  $1 \text{ AU} = 93.5 \times 10^6 \text{ mi} = 149.6 \times 10^6 \text{ km}$ .

When polar flattening is perceptible, both the polar and equatorial radii  $r_p, r_e$  can be determined or as in Table 3, the mean radius  $r = \frac{1}{2}(r_p + r_e)$  and the ellipticity  $\epsilon = (1 - r_p/r_e)$ . The mean radius may also be expressed in terms of the mean radius of the Earth (3959 mi) as a unit. The relative area is then very nearly equal to  $r^2$  and the relative volume to  $r^3$ .

**Masses, gravity, and density.** The mass of a planet is found easily if it has one or more satellites. If  $a$  is the mean distance (semimajor axis) of the satellite's orbit, and  $P$  its period of revolution

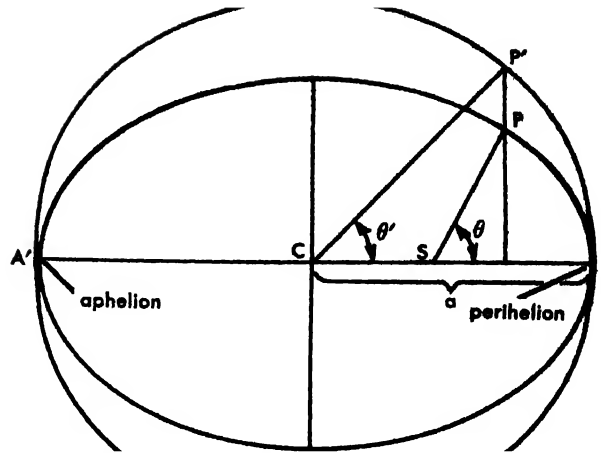


Fig. 4. Elliptic motion

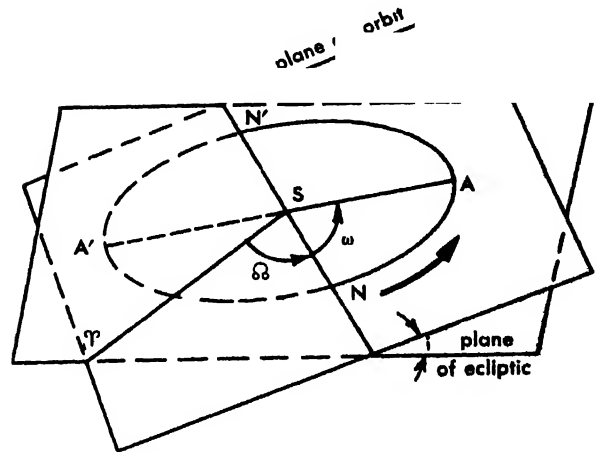


Fig. 5. Orbital elements.

expressed respectively in astronomical units and sidereal years, the mass  $m$  of a planet is given through Newton's law of gravitation by  $m = A^3/P^2$ , in terms of the mass of the Sun as a unit. This assumes that the mass of the satellite relative to the planet, and of the Earth relative to the Sun, may be neglected, which is nearly always the case within the accuracy of the data. Since the ratio

Table 2. Elements of planetary orbits

Planet	Symbol	Mean distance (semimajor axis)		Sidereal period of revolution		Synodic period	Mean velocity	Eccen- tricity	Inclination
		(AU)	( $10^6$ km)	(years)	(days)	(days)	(km/sec)	(epoch 1900)	(epoch 1900)
Mercury	☿	0.387	57.9	0.241	87.97	115.88	47.90	0.206	$7^{\circ}00'$
Venus	♀	0.723	108.2	0.615	224.70	583.92	35.05	0.007	$3^{\circ}24'$
Earth	♁	1.000	149.6	1.000	365.26		29.80	0.017	$0^{\circ}00'$
Mars	♂	1.524	227.9	1.881	686.98	779.94	24.14	0.093	$1^{\circ}51'$
Jupiter	♃	5.203	778.3	11.862	4332.59	398.88	13.06	0.048	$1^{\circ}18'$
Saturn	♄	9.546	1428.	29.458	10759.	378.09	9.65	0.056	$2^{\circ}30'$
Uranus	♅	19.20	2872.	84.018	30687.	369.66	6.80	0.047	$0^{\circ}46'$
Neptune	♆	30.09	4498.	164.78	60184.	367.49	5.43	0.009	$1^{\circ}47'$
Pluto	♇	39.5	5910.	248.4	90700.	366.74	4.74	0.247	$17^{\circ}09'$



Table 3. Physical elements of planets

Planet and symbol		Equatorial radius, $r_e$ , km ( $\delta = 1$ )		Ellip- ticity $\epsilon =$ $1 - r_p / r_e$	Volume ( $\delta = 1$ )	Mass ( $\delta = 1$ )	Den- sity, g/cm <sup>3</sup>	Escape velocity, km/sec	Rotation period	Tilt of axis <sup>a</sup>
Mercury	♿	0.38	2400	0.000	0.055	0.053	5.3	4.2	87.97d	small
Venus	♀	0.97	6200	0.000	0.91	0.815	4.95	10.3	weeks <sup>b</sup>	32° <sup>b</sup>
Earth	♁	1.00	6378	0.0034	1.00	1.000	5.52	11.2	23h56m4.1s	23.45°
Mars	♂	0.53	3400	0.0052	0.150	0.107	3.95	5.0	24h37m22.7s	25.0°
Jupiter	♃	11.20	71400	0.062	1317.	318.00	1.33	61.	9h50m30s <sup>b</sup>	3.1°
Saturn	♄	9.47	60400	0.096	762.	95.22	0.69	37.	10h14m <sup>c</sup>	26.73°
Uranus	♅	3.75: <sup>d</sup>	23800:	0.06	50.	14.55	1.56	22.	10h49m:	98.0°
Neptune	♆	3.50:	22300:	0.02	42.	17.23	2.27	25.	15h40m:	29°:
Pluto	♇	1 <sup>?</sup>	6400 <sup>?</sup>	<sup>?</sup>	1 <sup>?</sup>	0.9:	5 <sup>?</sup>	10. <sup>?</sup>	16h:	<sup>?</sup>

<sup>a</sup> To perpendicular to orbit.<sup>b</sup> Latitude <12° (system I); 9h55m40.6s, latitude >12° (system II).<sup>c</sup> Near Equator; 10h38m at intermediate latitudes.<sup>d</sup> Colons indicate uncertain determinations.

$m_r/M_s$  of the masses of the Earth and the Sun can be determined similarly, planetary masses are also known in terms of  $m_r$ .

If the planet has no satellite (Mercury, Venus, Pluto), its mass can be derived only from the perturbations it causes in the motions of the other planets and occasionally of comets. Since the perturbations are small, the masses so obtained are generally of low accuracy.

Once the mass  $m$  and the radius  $r$  of a planet are known in terms of the Earth's mass and radius, its surface gravity and mean density relative to the Earth are given by  $g = m/r^2$  and  $\rho = m/r^3$ , respectively. Multiplication by 981 and 5.552 gives the corresponding values in cgs units.

From  $r$  and  $m$  follows also the escape velocity  $V_1$  permitting a projectile (or a molecule) to leave the planet on a parabolic orbit:  $V_1 = (2Gm/r)^{1/2}$ ; this is  $\sqrt{2}$  times the velocity of an hypothetical satellite moving in a circular orbit close to the surface of the planet (see ESCAPE VELOCITY). These elements are listed in Table 3.

**Rotation periods.** The period of rotation of a planet is best determined by direct telescopic observation of the permanent markings on its surface (Mars, Mercury) or of the semipermanent cloud formations in its atmosphere (Jupiter, Saturn). When no definite details can be seen on the disk (Uranus, Neptune), the spectroscopic determination of the velocity difference between the opposite equatorial limbs can give, in combination with the linear diameter, an approximate value of the rotation period, unless it is so long (as is the case with Venus) that the shift of the spectral lines is not measurable. Finally, when the apparent diameter of the disk is too small for either of these two methods (Pluto, asteroids) a determination of the periodicity of the light variations, if any, due to the changing presentation of bright and dark regions of the surface may give a fairly accurate value of the rotation period. The rotation is now reasonably well known for all main planets except Venus, as indicated in Table 3.

**Planetary radiations.** The electromagnetic radiation received from a planet is made up of three main components: the visible reflected sunlight, plus some ultraviolet and near infrared radiation; the thermal radiation due to the planet's heat, including both infrared radiation and ultrashort radio waves; and the nonthermal radio emission due to electrical phenomena, if any, in the planet's atmosphere.

**Planetary brightness.** The apparent brightness of a planet can be measured by visual, photographic, and photoelectric photometry and is usually expressed in the stellar magnitude scale; it varies in inverse proportion to the squares of the distances  $r$  to the Sun and  $\Delta$  to the Earth. The fraction of the incident light reflected at full phase compared with the fraction that would be reflected under the same conditions by a perfect diffuse reflector is called the geometrical albedo. It is a measure of the reflecting power or reflectivity of the planet's surface. The visual albedos of the planets vary between 5 and 70%. See ALBEDO.

**Planetary atmospheres.** The chemical composition of a planetary atmosphere is derived from spectroscopic studies of the absorption bands, if any, present in the sunlight reflected by the planet. The major constituents of the atmospheres of the terrestrial planets are carbon dioxide, water, nitrogen, and (on Earth only) oxygen; of the major planets hydrogen, helium, methane, and ammonia.

**Heat radiation.** The heat radiation from a planet can be measured either with a radiometer at wavelengths of 8–15  $\mu$  (which are transmitted by the Earth's atmosphere) or with a radio telescope at wavelengths between 3 mm and 10 cm. In either case, the amount of energy corresponds to that which would be received under the same condition from a perfect radiator of the same size at a certain temperature  $T$ , called the black-body temperature of the planet (see HEAT RADIATION; RADIOMETRY). Its relation to the actual temperature depends on the properties of the atmosphere and surface of the planet. When the heat radiation is too weak to

measured (Uranus, Neptune, Pluto), the theoretical black-body temperature can be derived from the known value of the solar constant and the laws of heat radiation.

Nonthermal radio emission at meter wavelengths has been received from Jupiter by means of large radio telescopes. This emission takes the form of irregular bursts of noise originating in the planet's atmosphere from subagent sources, but its exact origin and mechanism are still unknown.

**Possible unknown planets.** During the nineteenth century, an unexplained irregularity in the motion of Mercury was thought by some to be caused by an unknown planet circulating between the Sun and Mercury, called Vulcan, which was looked for in vain. This irregularity was satisfactorily explained in 1915 by Einstein's general theory of relativity (see RELATIVITY). It is now certain that no intra-Mercurial planet of size comparable to the terrestrial planets exists.

The possibility of one or more planets circulating beyond the orbits of Neptune and Pluto has also been discussed, but there is no conclusive evidence for the existence of trans-Plutonian planets.

**Planets outside the solar system.** Minute perturbations in the elliptic motion of some nearby double stars have indicated the existence of minor components of small mass in these systems. The masses of these satellite bodies, although larger than planetary masses in the solar system, are considered to be too small to be self-luminous; they are consequently more like planets than dwarf stars. From this evidence it is inferred that planetary systems are not as uncommon in the universe as was previously believed, but no reliable estimate of the frequency of such systems can be made as yet. Whether some of the planets attending other stars are habitable by advanced beings or sustain some lower forms of life is unknown; the doctrine of the plurality of inhabited worlds remains a plausible, but unproved philosophical speculation.

For detailed information on the main planets, see JUPITER; MARS; MERCURY (PLANET); NEPTUNE; PLUTO; SATURN; URANUS; VENUS; see also ASTEROID; ASTRONOMICAL INSTRUMENTS; CELESTIAL MECHANICS; CERES; EARTH, ORBITAL MOTION; EROS; PERTURBATION (ASTRONOMY); RADIO ASTRONOMY; SOLAR SYSTEM; TROJAN PLANETS.

[G.D.V.]

**Bibliography:** H. N. Russell, R. S. Dugan, and J. Q. Stewart, *Astronomy*, vol. 1, rev. ed., 1945; W. M. Smart, *Celestial Mechanics*, 1953.

## Planetary

A projection device which faithfully portrays the heavens at any time in the past, present, or future. A planetarium is essentially a multiple slide projector, a typical one producing more than 110 separate images on an interior spherical projector screen or dome. At the opposite ends of a typical projection instrument are two large balls, each of which contains 16 projectors. These project onto

the white metal dome overhead 32 pictures which combine to form a representation of the principal stars in both hemispheres including, in the case of the largest planetariums, stars as faint as magnitude 6.5.

There are seven objects—the Sun, the Moon and the five inner planets—which are made to move against the background of the stars by means of paired projectors in the cages that separate the balls from the central portion of the instrument.

Motors provide motions for the planetarium. One set turns the instrument to simulate the daily rotation of Earth. Another set produces annual motion of the Sun, Moon, and planets forward or back in



Fig. 1. Charles Hayden Planetarium at Boston Museum of Science. (United Press International)

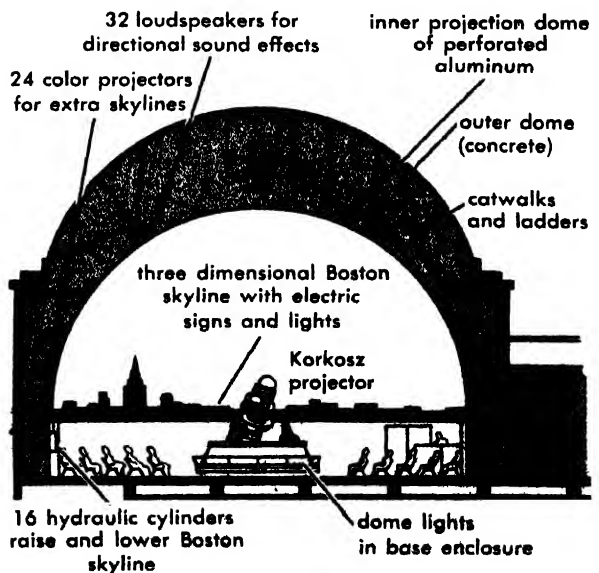


Fig. 2. Hemispherical planetarium screen produces illusion of open sky to audience. (Boston Museum of Science)

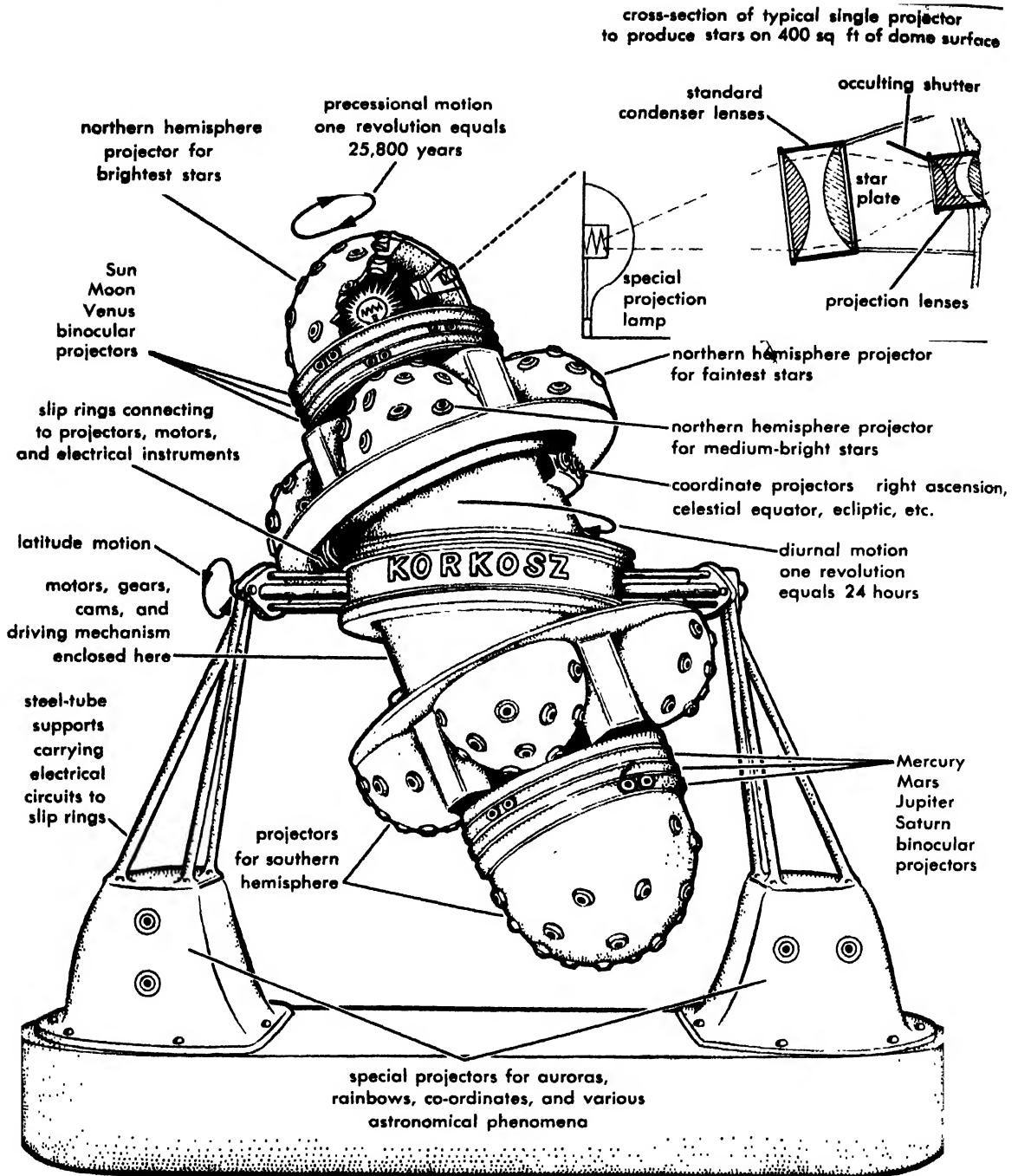


Fig. 3. Modern planetarium projects over 9600 stars. (Russell H. Long and *The Christian Science Monitor*)

time. Motion around a horizontal axis gives the effect of change of latitude. This permits the sky to be seen from any point on the surface of Earth from the North Pole to the South Pole. The motion of precession of the equinoxes, the wobbling motion of Earth's axis once in 25,800 years, can also be simulated, even to the extent of completing one 25,800-year cycle in 3 min. See PRECESSION OF EQUINOXES.

The planetarium chamber can be up to 75 ft in diameter and can seat up to 750 people (Fig. 2). A perforated hemispherical screen, which is the

dome, permits the projection of special effects during demonstrations. [I.M.L.]

### Planetary gear train

A sequence of meshed gears consisting of a central gear, a coaxial internal or ring gear, and one or more intermediate pinions, which are supported on a revolving carrier. In a simple planetary gear the pinions mesh simultaneously with the two coaxial gears (Fig. 1). With the central gear fixed, a pinion rotates about the central gear as a planet rotates about its sun; the gears are named accordingly.

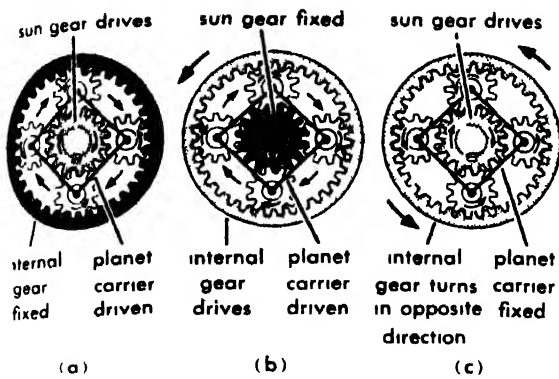


Fig 1 (a-c) Three principal modes of operation for simple gear train

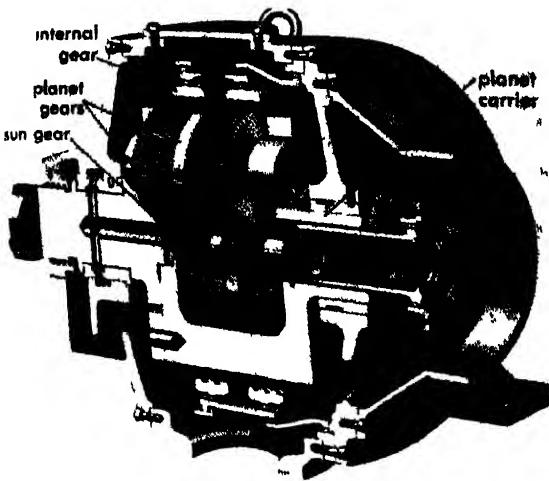


Fig 2 Stoeckicht planetary gear train. (DeLaval Steam Turbine Company)

Also with the sun gear fixed, a tooth on a planet pinion traces an epicycloid, hence the assembly is a form of epicyclic gear train. The planets and their carrier operate together and constitute a complete member. Figure 2 shows the typical construction of a planetary gear train.

In operation, input power drives one member of planetary gear train, a second member is driven to provide the output, and the third member is fixed. If it is not fixed, no power is transmitted through the gear train. This characteristic provides convenient clutch action; a brake band around the intermediate member and fixed to the gear housing serves to lock or free the third member without itself entering into the path of power transmission.

Any one of the three elements can be fixed: the central sun gear, the planet carrier, or the internal ring gear. Power can drive either of the two remaining elements and the other can deliver the output. There are thus six combinations of speed ratio with a single planetary gear train, with three being reciprocals respectively of the other three. The principal speed ratios for a simple gear train are, with internal gear fixed, large speed reduction (Fig. 1a)

$$\frac{\omega_b}{\omega_a} = 1 + \frac{N_c}{N_b}$$

with sun gear fixed, small speed reduction (Fig. 1b)

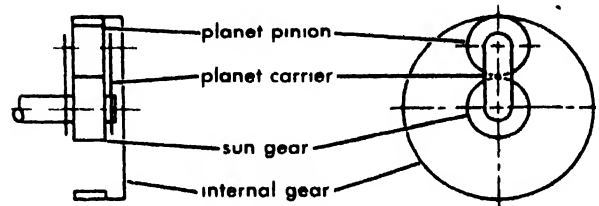
$$\frac{\omega_c}{\omega_a} = 1 + \frac{N_b}{N_c}$$

with planet carrier fixed, reverse, speed increase (Fig. 1c)

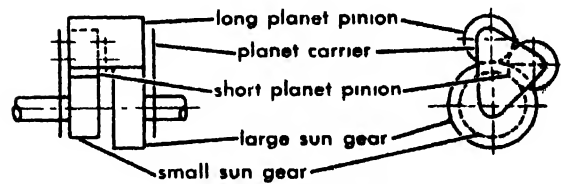
$$\frac{\omega_b}{\omega_c} = -\frac{N_c}{N_b}$$

where  $\omega$  is angular speed,  $N$  is number of teeth, and the subscripts identify the members:  $a$  for planet carrier,  $b$  for sun gear, and  $c$  for internal gear.

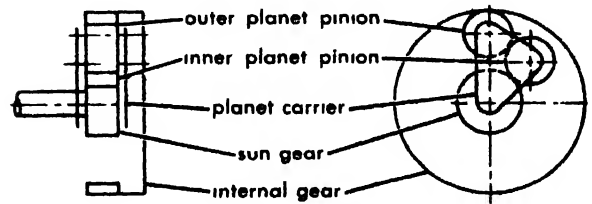
This speed-changing feature is used in automotive automatic transmissions. The number of teeth on the planet pinion of a simple planetary gear



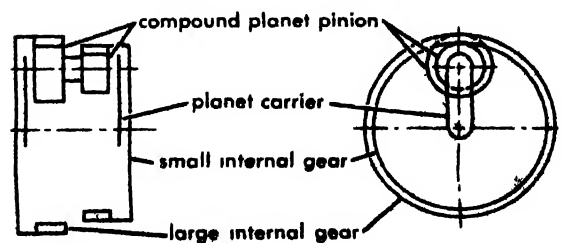
SIMPLE PLANET



EXTERNAL PLANET



DUAL PLANET



COMPOUND PLANET

Fig. 3. Four common arrangements of planetary gear trains.

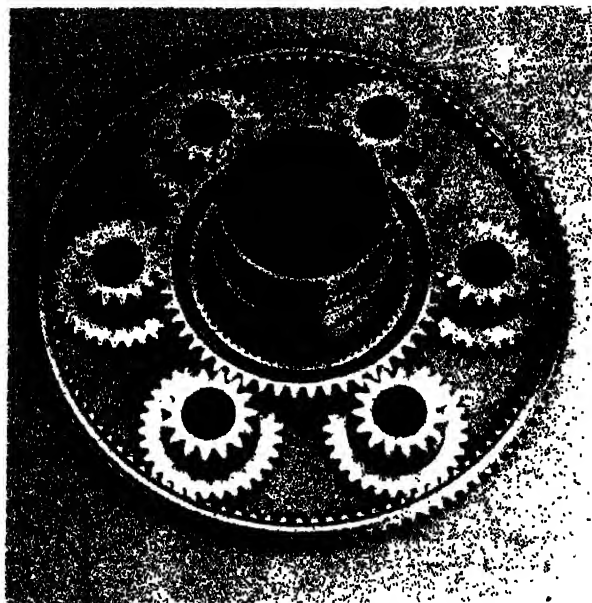


Fig. 4. Six compound planetary pinions transmit high power from aircraft engine to propeller with about 2:1 speed reduction. (Foote Brothers Gear and Machine Corp.)

does not enter into the equations for speed ratio because the pinion engages both sun and ring gears; hence the teeth of all gears are the same size, and the number of teeth on each gear is directly proportional to gear diameter. The above equations can, therefore, be expressed in terms of gear diameters instead of numbers of teeth.

Planetary gear trains can be variously modified for added flexibility (Fig. 3). Such gear trains provide large speed ratio in small space. With multiple pinions, they transmit large power (Fig. 4).

The aircraft engine drives the outer ring gear of Fig. 4; the central sun gear is fixed; the planet carrier arm, omitted in Fig. 4 to show the compound pinions, has stub shafts that carry the planet pinions in the manner of the cage of a ball bearing. The main shaft of the planet carrier passes through the opening in the sun gear to drive the propeller. The use of an internal gear accounts for much of the power capacity and smooth operation of the gear train. See GEAR TRAIN.

## Plant

A common term loosely used to designate any living organism not included in the animal kingdom. In the higher forms of life, a clear distinction can be made between plants and animals (Fig. 1). However, some of the lower organisms possess both plant and animal characteristics.

Although no single criterion can be used to separate all plants from all animals, certain features taken collectively provide a general basis for distinguishing between the two kingdoms of organisms. (1) Approximately five-sevenths of the 350,000 known plants have chlorophyll, a complex of green pigments which enables them, in the pres-

ence of light and air, to synthesize carbohydrate foods. Such plants are said to be independent, in contrast to bacteria, fungi, and nearly all animals—organisms that have no chlorophyll and are therefore dependent upon chlorophyll-containing plants for their supply of carbohydrates. (2) The embryonic tissues in most plants are abundant, persistent, and active, thus permitting an almost unlimited growth. Most animals, on the contrary, possess a limited scheme of growth because the embryonic or growth tissue is often used up in the process of maturation. (3) Almost all plants have a fairly firm structural framework of cellulose, a complex carbohydrate compound which is a major constituent of plant cell walls. Most animals, however, lack cellulose and their cells are almost universally enclosed in soft membranes rather than rigid walls. (4) The vast majority of animals have the power of locomotion, whereas only certain lower plants have the ability to move from place to place.

The lower forms of life that cannot be specifically classified as either plant or animal have been facetiously called "plantimals." T. H. Macbride

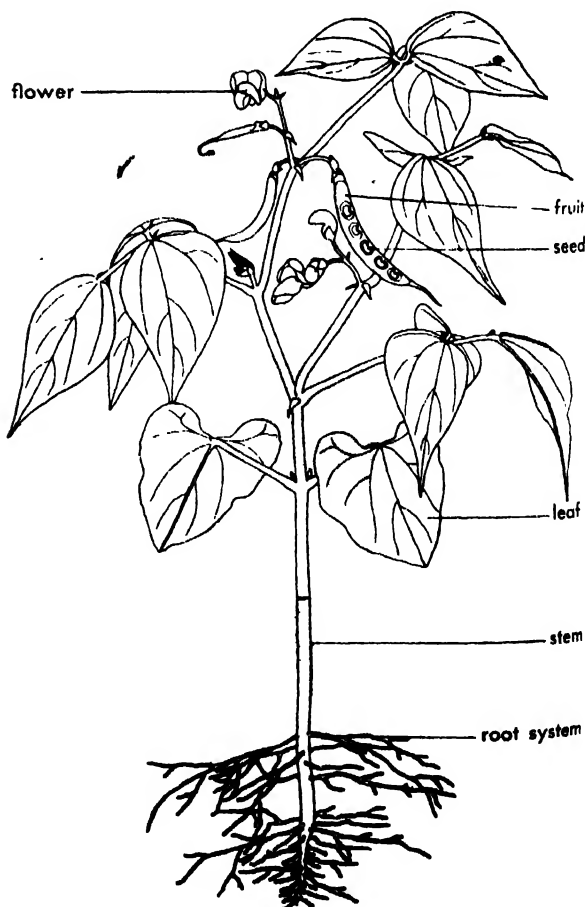


Fig. 1. Diagram of a complete herbaceous dicotyledonous plant (bean) showing the three reproductive organs (flower, fruit and seed), and the three vegetative organs (root, stem and leaf). (H. C. Sampson, *Workbook In General Botany*, Harper, 1949)

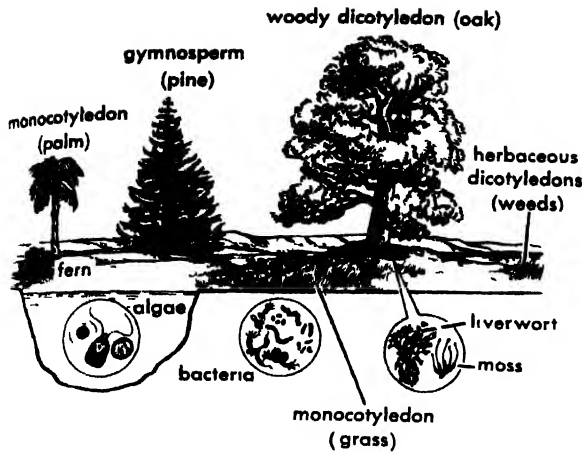


Fig 2 Plant forms.

and G. W. Martin, botanists, call the slime molds Myxomycetes, thus classifying them as plants; whereas A. Lister, a zoologist, calls them Mycetozoa, thus placing them in the category of animals. Likewise, certain mobile, one-celled, green organisms such as *Chlamydomonas*, *Volvox*, and *Pandorina* are regarded by the botanists as plants but are considered animals by zoologists. Most biologists agree, however, that plants and animals are fundamentally alike and have evolved from common ancestors during long periods of time.

Directly or indirectly, plants provide man and his domestic animals with all their food and with such other products as lumber, cork, textile fibers, cordage, vegetable oils, gums, resins, rubber, spices, perfumes, drugs, beverages, dyes, tanning materials, waxes, and paper.

Plants range from microscopic one-celled organisms such as bacteria and some algae to giant trees (Fig 2). For a classification of the varied plant forms see PLANT KINGDOM; see also PLANT ANATOMY, PLANT PHYSIOLOGY; PLANT TAXONOMY. [P.D.S.]

## Plant, mineral nutrition of

Beginning in the early 1800s, botanists became interested in determining the mineral requirements of plants. In order to determine the essential elements and the amounts of these that were required, scientists resorted to artificial culture techniques, and the plants were grown in water, sand, or some other inert medium in which they could be given a nutrient solution of known composition.

When all elements except one were supplied, it was possible to determine whether the omitted element was a required element. These experiments showed that plants require the following chemical elements: carbon, hydrogen, oxygen, nitrogen, potassium, phosphorus, sulfur, calcium, iron, magnesium, boron, manganese, zinc, copper, chlorine, and molybdenum.

Artificial culture techniques, often called hydroponics, proved useful and, in fact, mandatory for determining which elements were essential and which ones were not. This approach is still used in

an attempt to determine whether or not additional elements may also be required. The technique has been adopted, on a limited scale at least, for the commercial production of plants. In areas where there is no true soil, or in areas where the soil contains toxic constituents or pathogenic organisms, hydroponics has proved commercially practical.

When plants are grown hydroponically, arrangements must be made to provide their basic requirements: light, nutrient salts, water, aeration of the roots, and anchorage or support for the plants. Whether plants are grown in water culture, sand culture, or in a soil which has adequate nutrients, water, and aeration, plant growth and yields are identical. See HYDROPONICS.

**Essential elements.** Plants require various amounts of the essential elements. For elements such as potassium and calcium, a nutrient solution, or the soil solution of a good soil, may have as much as 100 parts per million (ppm). On the other hand, boron or manganese need be present only to the extent of  $\frac{1}{2}$  ppm, and, for molybdenum, 1 or 2 parts per billion. For micronutrient elements such as copper, boron, molybdenum, manganese, and zinc, the range of tolerance is quite narrow. Whereas  $\frac{1}{2}$  ppm of boron is adequate for most higher plants, as little as 2 ppm may prove lethal for certain species. See PLANT, MINERALS ESSENTIAL TO.

Fortunately, the concentration of most of the essential elements may vary over a considerable range without greatly altering plant growth and yield, so long as the concentration of a given essential element is not low enough to cause a deficiency or high enough to result in toxicity. Plant roots absorb relatively less of an element when it is present in a high concentration. This is undoubtedly an equalizing factor and explains why plants tend to grow equally well over quite a range of concentrations of the various essential elements.

Taking field corn plants as an example (Table 1), most of the dry weight is composed of carbon (43.6%), oxygen (44.4%), and hydrogen (6.2%). Carbon, as carbon dioxide, enters the plant through pores in the leaves called stomata. The carbon is built into carbohydrates, fats, proteins, and all of the other organic (carbon-containing) compounds of the plant. Amazingly, the plant obtains its vast amount of carbon from the air, in which there are only 3 parts of carbon dioxide per 10,000 parts of air.

Hydrogen and oxygen are derived from the water absorbed by the roots. The rest of the essential elements are absorbed from the soil. These latter elements are usually present in lower concentrations in plants than are carbon, oxygen, and hydrogen.

The nutritive elements are contained in the soil in two ways: dissolved in the water in the soil, or more or less loosely adsorbed onto and held by the surfaces of minute (less than 0.2 micron in



Table 1. Elements in corn plants\*

Element	Total dry weight of roots, stems, leaves, cobs, and grain, %
Oxygen	44.4
Carbon	43.6
Hydrogen	6.2
Nitrogen	1.5
Phosphorus	0.2
Potassium	0.9
Calcium	0.2
Magnesium	0.2
Sulfur	0.2
Iron	0.1
Silicon	1.2
Aluminum	0.1
Chlorine	0.1
Manganese	0.05
Undetermined elements	0.9

\* Adapted from W. L. Latshaw and E. C. Miller, Elemental composition of the corn plant, *J. Agr. Research*, 27:845-860, 1924.

diameter) soil particles called colloids. These colloids are of two types: mineral or clay colloids, and the organic colloids. The latter arise from the decomposition of organic matter. Whereas elements which are in solution may be leached from the root zone by rain, the elements which are adsorbed on the colloidal surfaces are resistant to loss by leaching. Per unit of surface area, organic colloids have a greater capacity than clay colloids to adsorb and to retain ions such as potassium, calcium, and magnesium (Fig. 1).

The most active portion of the root, with regard to salt absorption, is about  $\frac{1}{8}$  in. from the tip. The most active region for water absorption, on the other hand, is about  $\frac{1}{2}$  in. from the tip, where root hairs (extensions of epidermal cells) are most numerous. Inasmuch as these two regions tend to change and to mature with time, it is important for both salt and water absorption that the roots continue to grow. Otherwise the root "matures" almost to the tip, the root hairs in the root-hair zone die, and the root tips are no longer effective for either salt or water absorption.

**Salt absorption processes.** Elements may enter roots by any one or a combination of four processes. Three of these—diffusion, Donnan equilibrium, and ionic exchange or adsorption exchange—are purely physical processes. The fourth, active absorption or active transport, is a vital process which is dependent on the metabolism of living cells. For each of these types of ion entry, the first step involves adsorption of the ions on the surface of the roots.

**Diffusion.** Entry of ions by diffusion is a physical process in which ions enter the cells of the root only when the ions exist in a higher concentration outside the cells than on the inside. A given ion continues to enter only until the internal and external concentrations are equal. The cell plays no active role in the process, and ions enter or move solely by kinetic energy; that is, the ions are

in motion. This process cannot account for a concentrating of ions within cells, and there is no selection with regard to the types of ions which enter; nonessential and even toxic ions may enter.

**Donnan equilibrium.** This is a special type of diffusion which occurs when there are present, on one side of a membrane (that is, inside the cells), certain positively charged cations or negatively charged anions which are unable to pass through the membrane. It has been demonstrated by non-living, physical systems that this situation could lead to a higher concentration of diffusible ions inside root cells than in the external environment. Although root cells characteristically have higher concentrations of at least certain ions inside than outside the cells, the Donnan process is not regarded as an important one for the accumulation of ions by cells.

**Ionic exchange.** The process of ionic exchange, or adsorption exchange, is the third purely physical process by which ions may enter root cells. It is called ionic exchange because ions are exchanged in the process, as will be seen later, and adsorption exchange because it involves ions in an adsorbed state within the plant. This process may best be understood by thinking of the root as a sponge, the air pockets representing the vacuoles (cell sap) of cells and the solid portions of the sponge representing cytoplasm and cell walls. By means of this process, adsorbed ions remain outside of the vacuoles and are free to exchange position with ions in the external medium or in neighboring cells. The cytoplasm and cell walls (solid part of the sponge) are a continuum extending throughout the plant, and therefore the ions are free to move throughout the plant or to return to the external environment of the root. In this physical process, too, the plant plays no active role and

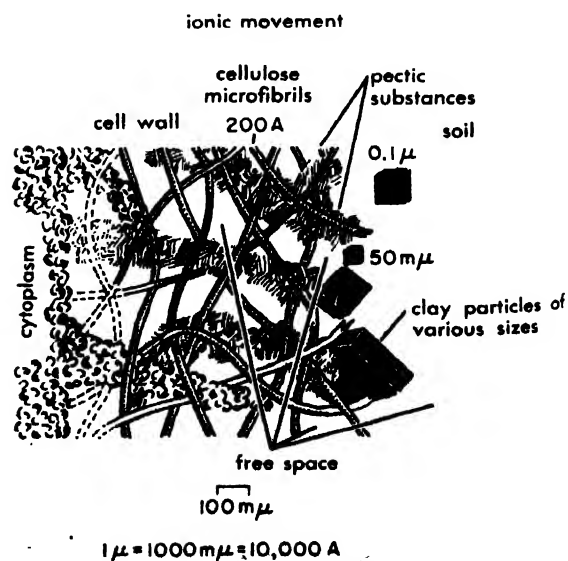


Fig. 1. Model of boundary region of root surface and growth medium. (From H. Jenny and J. Gonzalez, Modes of entry of strontium into plant roots, *Science*, 128:90-91, 1958)

shows no selectivity with regard to the entry of ions. Similarly, the process permits the concentration of an ion to become only as high internally as externally, and there is no build-up of ions within the root cells. For reasons that will become apparent when the next type of entry is discussed, ionic exchange is now regarded as the most important of the entry processes.

**Active absorption.** The fourth process by which ions enter roots is active absorption, or active transport. The term transport refers only to transport across a membrane and not to transport or movement throughout the plant. The term active indicates that the cells play a role in the process; only living, metabolizing cells are capable of this type of salt absorption. Cells must be carrying on respiration, the energy-liberating process (see BIOLOGICAL OXIDATION). Further, only aerobic respiration, involving oxygen, sustains this type of salt absorption.

If cells are deprived of oxygen and are therefore forced to carry on anaerobic respiration (in which oxygen is not involved and only one twenty-fifth as much energy is released), not only will cells fail to accumulate additional salt by this process, but they will lose to the external environment the ions of an element whose concentration is higher internally than externally. Under aerobic respiration, active absorption can result in a building up or concentrating of ions within the cells. This process is more important than the Donnan equilibrium, the other process which can result in a concentrating of ions in cells.

In the process of active absorption, the absorbed ions enter the vacuole or cell sap and, for the most part, tend to remain there. That is, their movement into the vacuole is largely a one-way, irreversible process. In the vacuole, therefore, the concentration of various ions may be many times the concentration of these same ions in the external solution. The cells must have a source of energy and do "work" in order to accumulate ions against such a concentration gradient.

Inasmuch as the ions which enter by this process pass irreversibly into the vacuoles (the air pockets of the sponge), they are not free to move to the top of the plant. For this reason, active salt absorption is not regarded as the important process by which ions enter and move to the top, where nutrient ions are also required. Active absorption is, however, the only salt absorption process in which there is selectivity with regard to the ions which enter. Thus this type of entry alone explains why different species of plants may have different internal concentrations of the various ions.

**Carriers.** The unique feature of active transport is that special protoplasmic constituents, "carriers," are believed to be involved. They are believed to be in the differentially permeable membranes of cells and to regulate the entry of cations and anions. Their chemical nature has not been determined, but they act much like the specialized, proteinaceous, organic catalysts called enzymes.

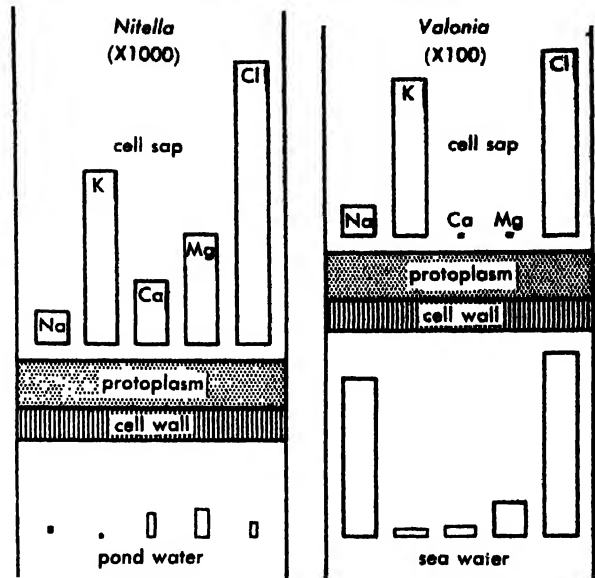


Fig. 2. Diagrammatic representation of the differences in absorption of ions by plants growing in low-salt and high-salt environments. (From P. J. Kramer, *Plant and Soil Water Relationships*, McGraw-Hill, 1949)

The kinetics of their action may be studied in the same manner as that of enzyme systems. It is postulated that they make a temporary union with an ion, that the resulting carrier-ion complex is unstable, and that, by a second reaction, the complex releases the ion on the other side of the membrane. The carriers are produced by the metabolism of cells, and they have a high degree of specificity with regard to the ions whose entries they regulate. For example, one carrier has been shown to be involved in the entry of potassium, rubidium, and cesium. Any one of these ions "competes" with the others for the cation-carrying "site" on this carrier. A different carrier is involved in the entry of sodium. The fact that potassium and sodium enter by the actions of different carriers may well explain why two species of plants differ so widely in their absorption of potassium and sodium from a common external environment. In other words, the two species of plants may differ widely in the concentrations of sodium- and potassium-carrying carriers. Therefore, differences in the permeability of roots to two kinds of ions may rest on the relative concentrations of the carriers that transport these ions. There are also specific carriers for lithium; for calcium, barium, and strontium; for magnesium; for sulfate and selenate; for chloride and bromide; for nitrate; for phosphate ( $\text{H}_2\text{PO}_4^-$ ), arsenate, and hydroxyl ( $\text{OH}^-$ ); and for phosphate ( $\text{HPO}_4^-$ ) and hydroxyl ( $\text{OH}^-$ ).

Under certain conditions or with time, during the life cycle of the plant, there are changes in the absorption rates of the various ions. Inasmuch as the carriers are formed by the metabolism of cells, and the carriers break down in time, the relative rates of production of the different kinds of

carriers may be altered by the cells as their metabolism changes. Therefore, not only do carriers explain the differences in rates of absorption of various ions by different species of plants, but changes in the relative concentrations of the carriers would explain changes in permeability with time.

Inasmuch as the carriers are involved in the irreversible movement of ions into the vacuoles of all cells, their presence and action would explain why the inorganic composition of roots, stems, and leaves of various species may differ markedly. From a common environment, one species might move considerable quantities of potassium into the vacuoles of cells and comparatively little sodium; another species might do just the opposite. Therefore, even though both species had the same concentrations of potassium and sodium in the cytoplasm and cell walls (by virtue of ionic or adsorption exchange), the differences in movements of ions into the vacuoles could explain the compositional differences of two species growing in a common environment.

Growing in dissimilar external solutions, two species of plants may show striking differences in their abilities to concentrate or to exclude certain ions. In Fig. 2 all ions shown reach a higher concentration in the sap of *Nitella* (a fresh-water plant) than in the external solution. In the case of *Valonia* (a plant that lives in sea water), the cell sap contains primarily potassium and chloride, whereas the sea water contains mainly sodium and chloride.

In view of the characterizations of ionic exchange and of active absorption, it is clear that the ions which are free to move to the tops of plants may be said to be the ones which escaped being absorbed (into the vacuoles) by the root cells. It is for this reason that ionic exchange is now regarded as the most important of the four types of salt absorption from the standpoint of plant nutrition. Differences in the inorganic composition of species are the result of active absorption, that is, the extent to which different ions are accumulated in the vacuoles of cells in the roots, stems, and leaves.

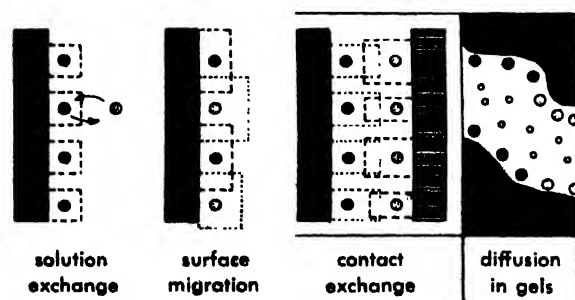


Fig. 3. Diagram of various modes of ionic migration, especially adsorption exchange. (From E. Truog, ed., *Mineral Nutrition of Plants*, Univ. of Wisconsin Press, 1951)

**Ion absorption from soil.** Ions, nutrient or non-nutrient ones, may be in aqueous solution in the soil solution or adsorbed onto the surfaces of colloidal particles. The colloidal surfaces of roots make intimate contact with the colloidal particles and, from the latter, the plant may take an adsorbed ion directly, without the ion's passing into solution. Nutrients which enter in this manner are said to do so by contact feeding or contact absorption. This is not, however, another type of salt absorption but rather refers to the external location of the ion, that is, whether it is in the soil solution or on the surface of a colloid. The first step in absorption from a colloid also involves adsorption of the ion onto the root surface. Actually, in contact feeding, the colloidal surfaces of the root and of the colloid are in such intimate contact, each with adsorbed ions, that an ion associated with the surface of the colloid may readily swap places with one associated with the root surface.

Whether an ion is absorbed from the soil solution or from the surface of a colloid, the plant must exchange an ion which it possesses for one which it absorbs (Fig. 3). Because the roots in carrying on respiration are producing carbon dioxide, the plant has a source from which to yield ions for ions it absorbs. Carbon dioxide dissolves in the water of the root cells to form carbonic acid which, in turn, dissociates into positively charged hydrogen ( $H^+$ ) ions and negatively charged bicarbonate ( $HCO_3^-$ ) ions. When a plant absorbs a cation, such as potassium, the root usually yields a hydrogen ion to the external medium.

The introduction of more and more hydrogen ions into the soil, in exchange for ions like potassium and calcium which are absorbed by the plant, explains why soils tend to become more acidic with time and why periodic liming is required. When the plant absorbs an anion, such as nitrate or phosphate, it most frequently yields a bicarbonate anion to the external medium (Fig. 4). The hydroxyl ( $OH^-$ ) anion may also be involved. It is interesting to note that the ions which the plant used in "swapping" for essential nutrients are hydrogen and bicarbonate ions produced as a result of respiration.

**Foliar nutrition of plants.** A plant may receive nutrients through the leaves as well as through the roots. Any of the essential elements may be supplied, at least in part, by application to the foliage. In fact, there may exist in the soil some condition which makes a certain element more or less unavailable to the roots. Iron and manganese, for example, may be rendered largely unavailable because of the alkalinity of certain soils. Under such conditions, the tops of the plants may show extreme iron or manganese deficiencies or both.

A foliar application of missing ions is a very practical way—indeed, the only way—of supplying these essential elements in such a situation.

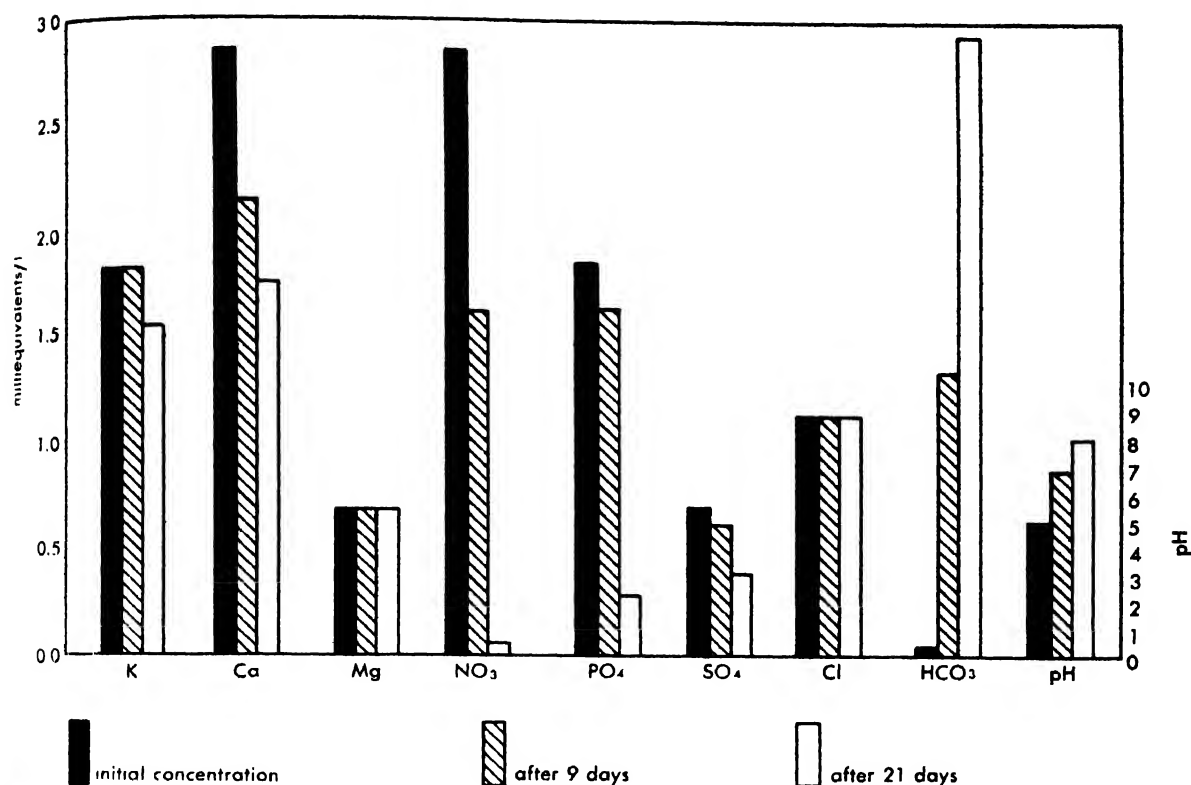


Fig. 4. Levels of principal ions initially in solution and after 9 and 21 days of a mass culture of the alga *Scenedesmus*. (From R. W. Krauss, *Inorganic nutrition*

of algae, in J. S. Burlew, ed., *Algal Culture from Laboratory to Pilot Plant*, Carnegie Inst. Wash. Publ. 600:85-102, 1953)

Often foliar nutrients are added to sprays that are being applied for the control of fungi, bacteria, or insects so that a separate spray for nutrients is not required.

Foliar nutrition may be a particularly useful approach during periods when the soil is waterlogged and the roots are relatively ineffective in the absorption of nutrients. Similarly, early in the spring, when the soil is cold, nitrification is proceeding slowly, and plant roots are relatively ineffective in absorbing nutrients, the application of nutrients—particularly nitrogen—to the tops may be most beneficial. Such conditions are particularly likely to occur with turfgrasses, such as on golf courses, and hence foliar nutrition is widely practiced. This approach is also widely used for supplying nitrogen, usually in the form of urea, to apple and citrus trees.

**Translocation.** For the most part, water and mineral salts absorbed by the roots move upward through the plant in the woody portion of the stem in a tissue called xylem. Organic substances such as sugars, which are made in the leaves, move downward to the roots through the bark of the stem in a tissue called phloem. However, sugars also use the phloem when they move upward from mature leaves of the plant to the growing point of the stem. Apparently all substances which come from the leaves move through the plant by way of the phloem.

One of the mysteries of transport or translocation which has long been studied is that some substances, such as sugars, may be moving in one direction in the phloem at the same time that other organic substances, such as amino acids, are moving in the opposite direction. Recent evidence indicates that some of the vertical strands of phloem cells are carrying materials in one direction while other strands of the phloem are carrying materials in the opposite direction. That is, an entire bundle or ring of phloem is not acting in unison.

There is no generally recognized explanation for the rapidity with which substances move in the phloem or for the amounts of substances which move through a given cross-sectional area of phloem in a given time. It is obvious that simple diffusion cannot begin to account for such rapid rates of movement of substances. Similarly, simple diffusion could not explain the vast quantities of sugars, for example, which have been shown to move through a given cross-sectional area of phloem to a rapidly developing fruit such as a pumpkin, squash, or watermelon.

The Münch mass flow hypothesis—which can account for movement of solution en masse in only one direction in the phloem—is not in accord with numerous reports of simultaneous, bidirectional movements in the phloem. Activated diffusion—movement of substances in or over interfaces—would account for the rapid, massive movements

of substances and even of substances moving in opposite directions. This type of movement is dependent solely on the relative concentrations of the substances in two regions of the plant. The concentrations of substances tend to equalize in the interfaces (such as planes of contact of cell wall with cytoplasm, or of portions of protoplasm with different densities of protoplasm), and this equalization process involves a rapid movement of materials. If a given substance, such as sugar being formed in the leaves, is added at one point, it spreads rapidly throughout the interfaces, moving toward regions with a lower concentration of sugar in the interfaces. Inasmuch as there may be a continual withdrawal of a substance from the interfaces, such as the removal of sugar by root cells, there tends to be a continual flow of sugar from the leaves, where an excess is produced, to the roots, which are dependent on the leaves for sugar.

Although activated diffusion is a purely physical process, active metabolism of living phloem cells is somehow required for the translocation of the organic substances in the phloem. It seems likely that the unique organization of protoplasm must be maintained in order for this physical process to occur, and cellular metabolism is undoubtedly required to maintain the unique physical structure of protoplasm. See CELL (BIOLOGICAL); PROTOPLASM.

It is generally recognized that the above-ground portions of the plant are dependent on the roots for water and mineral salts. Less generally recognized, however, is the dependence of the roots on the substances produced in the tops, particularly sugar. In perennial plants, renewed growth in the spring may be largely or entirely dependent on food reserves stored in the roots. For many of the plants whose tops die at the end of the growing season, renewed growth is entirely dependent on the roots. It is largely dependent on roots even for plants such as trees, whose trunks store some material for growth. This is shown strikingly when a complete ring of bark (in which the phloem is located) is removed from a tree by field mice. The tree will show few or no effects during the first season following the damage, but it may fail to grow the following spring, because destruction of the phloem ends translocation of food reserves to the roots. A valuable tree can often be saved by "bridge grafting" across such a break in the bark. See GRAFTING OF PLANTS.

**Relation of nutrition to yield.** The ultimate goal of commercial production of plants is yield—either of the vegetative plant, as in the case of hay, or of the fruit. Thus, the most practical and fundamental question about the mineral nutrition of plants is how to produce the greatest yield. It is obviously for this reason that man is interested in what chemical elements a plant requires and how much of each element it requires (Table 2).

Once it is known that all plants require potassium, for example, the problem then becomes one

Table 2. Raw materials used by corn plants\*

Substance	Approximate amount, lb/acre	Approximate equivalent
Water, H <sub>2</sub> O	4.3–5.5†	19–24 in. rain
Oxygen, O <sub>2</sub>	6,800	Air is 20% O <sub>2</sub>
Carbon, C	5,200 C, 19,000 CO <sub>2</sub>	Amount of C in 4 tons of coal
Nitrogen, N	160	Eight 100-lb bags of 20% fertilizer
Potassium, K	125	Three 100-lb bags of muriate of potash
Phosphorus, P	40	Four 100-lb bags of 20% superphosphate
Sulfur, S	75	78 lb of yellow sulfur
Magnesium, Mg	50	170 lb of epsom salt
Calcium, Ca	50	80 lb of limestone
Iron, Fe	2	2 lb of nails
Manganese, Mn	0.3	1 lb of potassium permanganate
Boron, B	0.06	¼ lb borax
Chlorine, Cl	Trace	Amount in rainfall
Zinc, Zn	Trace	Shell of one dry-cell battery
Copper, Cu	Trace	25 ft no. 9 copper wire
Molybdenum, Mo	Trace	2–3 oz ammonium molybdate

\* Figures are for corn plants producing 100 bushels per acre.

† In millions

SOURCE: G. Hambidge (ed.), *Hunger Signs in Crops*, 1st Symposium, Am Soc Agron and Natl Fertilizer Assoc, 2d ed., 1950

of determining how much potassium a given crop needs for maximal production. When there is insufficient potassium, the plant may show potassium deficiency symptoms, plant growth is stunted, and there may be little or no production of fruit. It would appear that there must be present in the plant some given concentration of potassium for maximal yield. With increasing amounts of potassium, over and above the level associated with deficiency, more and more growth occurs. Obviously beyond some given concentration of that element in the plant there is no further increase in yield with increase in internal concentration of that element. For some plants, the sufficiency values or levels for certain of the essential elements have been determined. This type of information is needed for all crop plants and for all of the essential elements if maximal production is to be attained.

If a crop is to produce a maximal yield, all factors, including the essential elements, must be at sufficiency concentrations or levels. If all of the essential elements are present in sufficiency except one, the growth and yield of the plants are limited to the extent that this essential element is lacking. Around 1840, the German scientist Justus von Liebig proposed this concept in the law of limiting factors. According to this law, growth is dictated by that element the concentration of which is the most insufficient. From this law it follows that, if the crop is only slightly deficient in phosphorus but extremely deficient in potassium, an increase in phosphorus will be essentially with-

out effect on growth and yield since potassium is the pace-setter. On the other hand, if potassium were added to bring the concentration of the element to sufficiency, growth of the plants would be increased only to the extent that the next most limiting factor, phosphorus, would permit.

Sometimes it appears obvious that a given crop is not growing as it should, although the plants do not indicate any particular deficiency. Upon the addition of a fertilizer containing nitrogen, phosphorus, and potassium, the plants may suddenly exhibit some clear-cut deficiency, in manganese, for instance. This may be explained on the basis that the plants were originally deficient, to varying degrees, in nitrogen, phosphorus, potassium and manganese. Upon the addition of the first three, clear-cut and striking manganese deficiency symptoms developed.

*Relation of inorganic to organic nutrients* The morganic nutrition of the plant cannot be evaluated without a consideration of its organic nutrition. Proteins, for example, contain carbon, hydrogen, oxygen, nitrogen, and sometimes phosphorus and sulfur. Protein formation depends on carbohydrates which are formed by photosynthesis. Therefore even though a plant were tested and found to contain less nitrogen than is known to be associated with maximal yield, the addition of nitrogen would be without beneficial effect if the plant were deficient in sugars because of prolonged cloudy weather or for some other reason.

*Soil moisture* Similarly, soil moisture is another factor in the mineral nutrition of plants. Absorption of nutrients by roots is related to the growth of roots. When soil moisture is inadequate, new root growth is limited or ceases, and salt absorption is accordingly decreased. From a practical point of view, then, there is no benefit in adding a known, deficient nutrient, such as nitrogen, if the roots are not prepared to absorb it. Nutrient salts cannot substitute for water, nor can water substitute for nutrients. The interdependence or interrelationships between water, nutrients, sugars, and so on are often overlooked. Successful agricultural experts in the field of crop production have to take all of these factors into consideration. It is important economically to know what factor is limiting plant growth, even though it is a factor, such as sunshine, over which man has no control. There would be no advantage in adding a fertilizer if sunshine were the limiting factor, and hence the cost of a useless application of fertilizer could be saved. On the other hand, if some essential element were found to be the limiting factor, its addition might more than pay for the cost of the added fertilizer. See FERTILIZER.

The most frequent limiting factor in crop production is undoubtedly the amount of soil moisture. Only in recent years has this become generally recognized, the result being the installation of facilities for supplemental irrigation even in relatively humid regions of the United States, such as

the eastern states. In many cases, the installation of an irrigation system has paid for itself in a single year, when, because of a prolonged drought, the yield of the crop would otherwise have been a total failure. See SOIL.

**Excessive salt concentration toxicity.** In the semiarid or arid regions of the West, where irrigation is widely practiced, the low rainfall and the salt content of the irrigation waters combine to cause undesirably high concentrations of salts in many of the soils. The low rainfall is insufficient to leach accumulated salts from the soil, and the irrigation waters naturally contain appreciably more salt than does rain water. For example, the water of the Colorado River, which is widely used in irrigation, contains about 700 ppm of salt. About 27,000,000 acres of land in the Midwest and Far West are under irrigation. Even in areas which are not already clearly "salinity" areas, there is always a potential danger of the development of salinity. In some of these soils, the salts which accumulate are alkaline in reaction, and the soils are called alkali soils. When the accumulated salts are essentially neutral in reaction, such as sodium chloride or sodium sulfate, the soils are properly called saline soils. Salinity is regarded as a problem when the concentration of salts reaches 0.2% of the dry weight of a loam soil; most species of plants die when the concentration is around 2%.

Salinity conditions are often unsuspected, inasmuch as there are no plants of the same species growing on nonsaline soil in the same area. Therefore, actual reductions in growth and yield may go undetected if these are not too pronounced. With higher levels of salinity, it is obvious that plant growth is restricted, and the plants may show "burning" or "firing" of the leaves, particularly the oldest, lower leaves, which have had the most time to accumulate toxic concentrations of salts.

In contrast with deficiency symptoms, the predominance of a given toxic salt in the soil does not usually manifest itself in symptoms indicative of the type of salt which is present. On the basis of symptoms, one cannot ordinarily determine whether there is a high concentration of sodium chloride or of sodium sulfate in the soil. The symptoms are very much the same--stunting and firing of the leaves, particularly at the tips, margins, or both.

A toxic concentration of boron in the soil solution induces rather specific symptoms, particularly on the oldest leaves, which properly trained scientists can usually identify.

Soil analyses, of course, may be used to determine the nature of the accumulated salts in saline or alkali soils. Often this information is useful in prescribing a remedy for the removal of salts. If the soil is permeable enough to water and if sufficient irrigation water (preferably of low salt content) is available, the toxic salts may be leached out of the root zone and carried away in the drainage water.



When sodium salts accumulate, they usually impart undesirable physical characteristics to the soil, such as reduced permeability to water and reduced aeration. If the sodium concentration is not too high and if the value of the land is sufficient to warrant it, this condition can often be overcome by the application of a calcium-containing salt, such as gypsum, to the soil. Calcium replaces the sodium on the clay and organic colloids, and the sodium may then be leached from the root zone. With the substitution of calcium for sodium, the permeability of the soil to water usually improves with time so that it becomes increasingly easier to wash the sodium salts from the root zone. Along with this change, the over-all physical characteristics of the soil, including tilth and aeration, become more conducive to plant growth. See PLANT METABOLISM. [H.G.G.]

**Bibliography:** H. G. Gauch, Mineral nutrition of plants, *Annual Review of Plant Physiology*, 8:31-64, 1957; D. R. Hoagland, *Lectures on the Inorganic Nutrition of Plants*, 1944; B. S. Meyer and D. B. Anderson, *Plant Physiology*, 2d ed., 1952; W. Reuther (ed.), *Plant Analysis and Fertilizer Problems*, American Institute of Biological Sciences, Publication no. 8, 1961; L. A. Richards (ed.), *Diagnosis and Improvement of Saline and Alkali Soils*, U.S. Department of Agriculture, Handbook 60, 1954; E. Truog (ed.), *Mineral Nutrition of Plants*, 1951.

## Plant, minerals essential to

Beginning in the middle of the nineteenth century, botanists sought to determine the chemical elements required for the growth of plants. Elements such as nitrogen, phosphorus, and potassium, required in relatively large amounts, were among the first shown to be essential. The essentiality of certain elements, such as copper and molybdenum, was not established until chemical compounds of greater purity were produced. Earlier, many of the elements were present in sufficient amounts as impurities in a nutrient solution to preclude their detection as essential elements when they were "omitted" from the solution. Elements which are required in small amounts, such as copper, molybdenum, manganese, zinc, boron, iron, and chlorine, are sometimes called trace elements. They have also been called minor elements, but this term erroneously implies that these elements play only a minor role in plant nutrition (see PLANT, MINERAL NUTRITION OF). The term micronutrients is now generally favored, since the term implies that small amounts are required and that they perform a nutritive role.

At the present time, most plant scientists agree that the following elements are essential for higher plants: carbon, hydrogen, oxygen, phosphorus, potassium, nitrogen, sulfur, calcium, magnesium, iron, boron, manganese, zinc, copper, chlorine, and molybdenum. The last seven are micronutrients. Nitrogen, phosphorus, and potassium are the three

most important and are the components of a 5-10-5 fertilizer (5% nitrogen, 10% phosphorus as phosphorus pentoxide, and 5% potassium as potassium oxide). Certain algae have been shown to require vanadium, sodium, and cobalt. There is a good possibility that additional elements may be added, from time to time, to the present list of essential elements for plants.

**Criteria of essentiality.** In order to be considered essential, an element must meet the following criteria: (1) absence of the element must result in abnormal growth, injury, or death of the plant; (2) the plant must be unable to complete its life cycle without the element; (3) the element must be required for plants in general; and (4) no other element must be able to serve as a complete substitute. Most scientists would add a fifth criterion, namely, that the element must be shown to have a specific and direct role in the nutrition of the plant. This last criterion is the most difficult of all to determine since known, direct roles for potassium and calcium, for example, are not as yet agreed upon. Their essentiality, however, is unquestioned. Without exception, the essential elements were first shown to be so when their removal from the external environment caused drastically reduced growth. The direct, specific roles of the elements were then pursued, and most of these have been elucidated.

**Roles of the essential elements in plants.** The following paragraphs discuss the roles in plants of carbon, hydrogen, oxygen, phosphorus, nitrogen, sulfur, potassium, calcium, magnesium, iron, manganese, copper, zinc, molybdenum, boron, and chlorine.

**Carbon.** Although not a mineral, carbon is an essential element and may constitute as much as 44% of the dry weight of a typical plant such as corn. It is a component of all organic compounds such as sugars, starch, proteins, and fats, in plants (and animals). One of its main roles, then, is that of being a constituent of these important compounds and of the vast array of compounds which, in turn, are synthesized from sugar. Carbon may indeed be said to be involved in all of the roles played by all of the carbon-containing compounds. This would include such compounds as the plant hormones, which regulate plant growth, flowering, and reproduction. See CARBOHYDRATE; PLANT HORMONES; PLANT METABOLISM; PLANT MORPHOGENESIS; PROTEIN; STARCH.

**Hydrogen.** Along with carbon, this nonmineral element is also a constituent of all organic compounds. All that has been said of carbon would thus similarly apply to hydrogen. Hydrogen constitutes about 6% of the dry weight of a plant.

**Oxygen.** This nonmineral element is a constituent of all carbohydrates, fats, and proteins and, in fact, of most organic compounds. Some organic compounds, such as carotene, which gives rise to vitamin A, are composed only of carbon and hydrogen. Inasmuch as oxygen is a constituent of most organic compounds, like carbon and hydrogen, it

may be said to be involved in whatever functions those compounds perform. Oxygen constitutes about 43% of the dry weight of a plant.

Carbon, hydrogen, and oxygen are constituents of the bicarbonate ion ( $\text{HCO}_3^-$ ) which is believed to be one of the chief anions (along with hydroxyl,  $\text{OH}^-$ ) exchanged by the plant for anions absorbed from the soil. Because of this role, these three elements may be said to be involved in the process of salt absorption by roots.

Finally, in the process of oxidation (aerobic respiration) of foods by the cell, oxygen is the final acceptor of hydrogen. When a food, such as sugar, is respired by the removal of hydrogen and carbon dioxide, the latter and water appear as end products. Water is formed in the terminal step in these oxidation reactions when oxygen and hydrogen unite. See BIOLOGICAL OXIDATION.

**Phosphorus.** Certain special proteins in all cells, the nucleoproteins, contain phosphorus. As the name implies, these special proteins are in the nucleus of the cell and hence in the chromosomes which carry the hereditary units, the genes. Inasmuch as phosphorus is a constituent of these proteins of the nucleus, cell division may be said to be dependent on phosphorus. The transmission of hereditary characteristics also depends on this element. See CHROMOSOME; NUCLEOPROTEIN.

Cellular membranes are believed to consist, in part, of special phosphorus-containing fats or lipids called phospholipids (see PHOSPHATIDE). These, along with hydrated protein molecules, are very likely the chief components of cellular membranes. Differentially permeable membranes regulate the entry and exit of materials from the cells, and so phosphorus may be said to play a role in the permeability of cells to various substances and in the retention of substances by cells.

Phosphorus is a constituent of some special compounds, diphosphopyridine nucleotide (DPN) and triphosphopyridine nucleotide (TPN). These unique compounds are involved in the transfer of hydrogen in aerobic respiration, and life itself depends on this all-important energy-liberating process. See DIPHOSPHOPYRIDINE NUCLEOTIDE (DPN); TRIPHOSPHOPYRIDINE NUCLEOTIDE (TPN).

The early stages in the combustion or utilization of sugar by cells involve the addition of phosphorus to the sugar molecule. Only after phosphorus is added to both ends of the sugar molecule is it cleaved and prepared for further transformations which release the chemical energy stored in the sugar molecule.

Finally, phosphorus is a constituent of unique compounds called adenosinediphosphate (ADP) and adenosinetriphosphate (ATP). In the former compound, one of the two phosphate bonds is a high-energy bond, and in the latter, two of the three phosphate bonds are high-energy bonds. The unique feature of these compounds is the concentration of energy in these special phosphorus bonds. This special bond is one of the ways in which potential energy is stored within the cell. The bond

is important not only because it represents a form of stored energy, but also because such energy may be used to accomplish "work" when the bond is broken and the energy is released. Many synthetic reactions, such as the syntheses of sucrose and starch from glucose, require energy which the high-energy phosphate bond is capable of delivering to the reactions. See ADENOSINEDIPHOSPHATE (ADP); ADENOSINETRIPHOSPHATE (ATP).

The reduced forms of diphosphopyridine nucleotide ( $\text{DPNH}_2$ ) and triphosphopyridine nucleotide ( $\text{TPNH}_2$ ) constitute the other major form of energy storage in cells. This chemically stored energy is also available for work—driving certain chemical reactions that would otherwise proceed at imperceptibly low rates.

**Nitrogen.** Along with carbon, hydrogen, and oxygen, nitrogen is a constituent of all amino acids, the building blocks of the proteins. Protoplasm usually has a high percentage of water, but the substance portion is primarily proteinaceous. Thus nitrogen and certain other elements such as carbon, hydrogen, and oxygen may be said to be a part of the living substance, protoplasm.

All enzymes thus far isolated have been shown to be protein in nature. Therefore, nitrogen is one of the constituents of these remarkable organic catalysts which can accomplish at room temperature, or below, chemical reactions which man can perform only with high temperature, pressure, or other special conditions.

Nitrogen is also a constituent of the chlorophyll molecule, there being four nitrogen atoms in each molecule. Nitrogen is thus directly required in photosynthesis, a food-manufacturing process which only plants can accomplish. See CHLOROPHYLL.

**Sulfur.** Certain amino acids, such as cystine and cysteine, contain sulfur. These sulfur-containing amino acids are often components of plant proteins and, less frequently, of animal proteins. Sulfur is also a constituent of the tripeptide glutathione, a pigment which may function as a hydrogen carrier in the respiration of plants as well as in animals. See CYSTEINE; CYSTINE.

**Potassium.** Although potassium was one of the first elements shown to be essential and is required in relatively large amounts by plants (often 1% or more of the dry weight), there is no known compound in plants which contains potassium. Despite numerous researches, it still is not known why plants require such seemingly large amounts of the element. It may function as a cofactor for certain enzyme systems, but it would not appear that such high concentrations of potassium should be required for this purpose. Virtually all of the potassium in plants appears to be water-soluble, a point which further emphasizes the fact that potassium is not a constituent of any compound and certainly not of the larger, relatively insoluble and immobile compounds in plants. See ENZYME.

**Calcium.** Although calcium may typically be present in plants to the extent of 0.2% of the dry

matter, it also is not clear why plants require so much calcium. Many workers consider the cementing substance between cells, the middle lamella, to be composed of calcium pectate. No other calcium-containing compounds of biological significance have been reported for plants, and yet they contain far more calcium than would be required for calcium's postulated role in the middle lamella. Excesses of oxalic and other organic acids may appear in the cell as crystalline calcium salts of low solubilities. These salts, however, are considered waste products and serve no useful, vital function. Their removal from solution by calcium may prevent a toxicity that would otherwise result from such acids.

**Magnesium.** One of the earliest roles discovered for magnesium was as a constituent of the chlorophyll molecule. Each molecule contains one atom of magnesium in the center. Although other metallic ions may be made to replace magnesium in the chlorophyll molecule, chlorophyll functions in photosynthesis only when it contains a magnesium atom. Despite much speculation, no one has yet determined why only magnesium is effective in chlorophyll and in photosynthesis.

In addition to the unique role which magnesium plays in chlorophyll and photosynthesis, it is also required for the action of a host of enzymes. It is also apparently required for an enzyme concerned with oil formation in plants, since oil droplets are not formed in the alga *Vaucheria* in the absence of magnesium. Also, the seeds of plants, which contain large amounts of oil, are consistently high in magnesium.

**Iron.** Iron is required for the formation of chlorophyll but is not a constituent of the molecule. In leaves, about 80% of the iron is associated with chloroplasts, the chlorophyll-containing plastids. Iron is a constituent of cytochrome f, which may have a unique role in photosynthesis.

The element is a constituent of the enzymes cytochrome oxidase, peroxidase, and catalase. The first of these is involved in respiration and the last catalyzes the breakdown of any hydrogen peroxide that forms in cells as a result of certain metabolic reactions. As with calcium, potassium, and magnesium, there is no evidence as to why plants require as much iron as they do, since much less iron would appear to suffice for its known roles.

**Manganese.** There are no known compounds in plants of which manganese is a constituent. There is considerable evidence that the element may be a cofactor or an activator in certain enzyme systems. For example, manganese may be involved in nitrate reduction and hence in nitrogen metabolism. Manganese has been clearly shown to be required for photosynthesis, most strikingly by algae growing in an inorganic medium. See PHOTOSYNTHESIS.

**Copper.** Copper is a constituent of the enzymes laccase, ascorbic acid oxidase, and tyrosinase (polyphenoloxidase). The last enzyme is believed to be involved, in most plants, with the terminal step in aerobic respiration, the transfer of hydro-

gen to oxygen to form water. This action thus links copper with energy release in plant cells.

**Zinc.** Although the enzyme has not been isolated, it has been shown that the enzyme which synthesizes the amino acid tryptophan requires zinc. Tryptophan, in turn, is the precursor from which the plant hormone indoleacetic acid is made. It may thus be said that zinc is directly necessary for the formation of tryptophan and indirectly necessary for the production of indoleacetic acid. See TRYPTOPHAN.

Two zinc-containing enzymes have been isolated from plants, namely, carbonic anhydrase and alcohol dehydrogenase.

**Molybdenum.** Molybdenum is required, in the external solution, to the extent of 1 or 2 parts per billion. In the dry matter of the plant, it may be present only to the extent of around 10 parts per billion. It has been calculated that the number of molybdenum atoms required per cell of *Scenedesmus obliquus* and *Azotobacter* is 3,000 and 10,000, respectively.

Molybdenum is the metal component of the enzyme nitrate reductase, which effects the reduction of nitrate nitrogen to the reduced form of nitrogen which is incorporated into amino acids and then proteins. Molybdenum-deficient tomato plants may accumulate nitrate to the extent of 12% of the dry weight of the plant. If such plants are given a few parts per billion of molybdenum in the external medium, the nitrate content will drop to around 1% within 2 days. *Aspergillus niger*, *Scenedesmus obliquus*, and *Chlorella pyrenoidosa* also require molybdenum for the reduction of nitrate nitrogen. Fixation of atmospheric nitrogen by one of the free-living, nitrogen-fixing bacteria, *Azotobacter*, and by a blue-green alga, *Anabena cylindrica*, requires molybdenum. The element is, therefore, intimately associated with nitrogen metabolism, synthesis of protein, and, hence, synthesis of protoplasm. See NITROGEN CYCLE.

Certain species of plants appear to require molybdenum for one or more unidentified roles other than nitrate reduction or nitrogen fixation. For example, cauliflower plants grown on urea and ammonium, as reduced nitrogen sources, nevertheless develop characteristic molybdenum deficiency symptoms known as whip tail when molybdenum is withheld.

**Boron.** The essentiality of this element was first established around 1910. Approximately  $\frac{1}{2}$  part per million (ppm) in the external solution suffices for the growth of most plants. Garden and sugar beets, as well as alfalfa, have a somewhat higher boron requirement, 5–10 ppm being optimal.

There are no known compounds in plants of which boron is a constituent, and no enzyme system has been shown to require boron. In most plants, boron is immobile, suggesting that it is combined with large, immobile molecules. Owing to the immobility of boron, plants have to receive boron continually throughout their life cycles.

Numerous roles have been proposed for boron, including roles in carbohydrate and protein metabolism. One of the more recent theories is that boron is required for the translocation of sugar from the leaves, where sugar is made, to the flowers, fruits, and the growing points of stems and roots. In the absence of boron, stem and root tips die and flowering and fruiting are drastically reduced or altogether curtailed. A given degree of deficiency of boron, which results in almost complete failure to set seeds in alfalfa, may not materially reduce the size of the plants. This well-established phenomenon and others signify a unique role of boron in flowering and fruiting. Successful germination of pollen grains and the production of the pollen tubes require boron.

Boron-deficient plants lose their normal response to gravity, indicating that boron is involved in the production, movement, or action of the natural plant hormones that cause the stem of a horizontally placed plant to turn up and the roots to turn down.

**Chlorine.** For a plant such as tomato, a deficiency of chlorine results in a wilting of the leaf tips and chlorosis (yellowing), bronzing, and necrosis (death) of the leaves. If chlorine is added early enough, as little as 3 ppm banishes the symptoms and normal growth proceeds.

Tomato plants show chlorine deficiency when they contain around 250 ppm of chlorine (dry weight basis), whereas this species shows molybdenum deficiency only when the concentration is around 0.1 ppm. Therefore, the tomato plant requires several thousand times as much chlorine as molybdenum.

It should be made clear at this point that plants cannot tolerate more than a few parts per million of chlorine, that is, the molecular, gaseous form of the element. Ordinarily plants absorb the element in the ionic form, that is, as chloride. Most plants can tolerate 500 ppm or more of chloride without much affecting growth, and certain halophytes (salt plants) can grow vigorously in high concentrations of chloride salts.

**Other elements required by certain plants.** Vanadium is required for the growth of *Scenedesmus obliquus*, and it has been shown to play a role in photosynthesis in *Chlorella*. There is no evidence of its essentiality for plants other than the green algae.

Sodium is an essential element for certain blue-green algae, but according to the criteria of essentiality, it is not required for green algae or higher plants.

Cobalt is required only for certain blue-green algae.

**Deficiency symptoms.** A deficiency of any one of the 16 essential elements results in stunted growth and reduced yield.

Deficiency symptoms are best identified by persons specifically trained to recognize them, since a deficiency of a given element appears quite different on different plants, such as corn and beans.



Fig. 1. Young tobacco seedling showing potassium deficiency symptoms consisting of interveinal chlorosis and marginal and tip "scorch" of older leaves. (From G. Hambridge, ed., *Hunger Signs in Crops; A Symposium*, Am. Soc. Agron. and Natl. Fertilizer Assoc., 2d ed., 1950)

Furthermore, the application of nutrients to correct deficiencies of most of the elements, particularly boron, copper, manganese, zinc, and molybdenum, calls for the knowledge of specialists in plant nutrition.

The elements which are most likely to have a limiting effect on growth are nitrogen, phosphorus, and potassium; these are present in a typical, readily available fertilizer such as 5-10-5. Generalizations can be made with regard to the deficiency symptoms of these three main elements. Inasmuch as nitrogen deficiency results when nitrogen moves out of the older, hence lower, leaves of a plant, this deficiency is generally characterized by a yellowing of these leaves. Phosphorus deficiency is often characterized by a purpling of the stem, of the leaf, or of veins on the underside of the leaves.

In corn, phosphorus deficiency causes a purpling of the stem and, at times, a purpling of the leaf blades. Potash (potassium) deficiency results in a burn or scorch of the margins of the leaves, particularly the older, lower leaves (Fig. 1). Recognition, then, of the deficiency symptoms of these three elements is important and desirable, since any one or more of these deficiencies can be corrected by the application of readily available commercial fertilizer.

Chemical tests ("tissue tests") can often be made of key plant tissues to determine whether a given element is lacking. Such tissue tests have the advantage of detecting a near-deficiency before it becomes acute enough to express itself in the form of deficiency symptoms. In general, such tests must be made by persons trained to conduct and to interpret them.

The best approach for the average homeowner or farmer, however, is to have the soil tested, if there is any question as to its productive capacity. There are commercial laboratories which provide this service as well as the state agricultural experiment stations. A soil test has the advantage of predicting, in advance of planting, what nutrients are lacking. By the time deficiencies appear, plant



Fig. 2. Boron deficiency in grape plant showing interveinal chlorosis of terminal leaves and necrotic terminal growing point. (From J. A. Cook et al., *Light fruit set and leaf injury from boron deficiency in vineyards readily corrected when identified*, Calif. Agr., 15(3):3-4, 1961)

growth and yield are usually irretrievably retarded. Tissue tests, if they are used early enough, can detect an incipient deficiency for correction.

In addition to the already described, widespread need for nitrogen, phosphorus, and potassium, it is often necessary to add certain other elements. The following elements have been found to be deficient in one or more areas of the United States: boron (Fig. 2), magnesium, copper, manganese, zinc, iron, calcium, sulfur, and molybdenum. A deficiency of chlorine has not been observed under field conditions. In one soil or another, a deficiency of every essential element except chlorine has been found in nature. Considering the number of years that some of the soils have been in cultivation and the amounts of the essential elements which have been removed by the crops, it is not surprising that agricultural soils are becoming deficient in more and more of the essential elements. [H.G.G.]

**Bibliography:** G. Hambidge (ed.), *Hunger Signs in Crops; A Symposium*, Am. Soc. Agron. and Natl. Fertilizer Assoc., 2d ed., 1950; C. A. Lamb, O. G. Bentley, and J. M. Beattie (eds.), *Trace Elements*, 1958; T. Wallace, *Trace Elements in Plant Physiology*, 1950.

## Plant, water relations of

Water is the most abundant constituent of all physiologically active plant cells. Leaves, for example, have water contents which lie mostly within a range of 55-85% of their fresh weight. Other relatively

succulent parts of plants contain approximately the same proportion of water, and even such largely nonliving tissues as wood may be 30-60% water on a fresh weight basis. The smallest water contents in living parts of plants occur mostly in dormant structures such as mature seeds and spores. The great bulk of the water in any plant constitutes a unit system. This water is not in a static condition. Rather it is part of a hydrodynamic system which in terrestrial plants involves absorption of water from the soil, its translocation throughout the plant, and its loss to the environment, principally in the process known as transpiration.

**Cellular water relations.** The typical mature vacuolate plant cell constitutes a tiny osmotic system, and this idea is central to any concept of cellular water dynamics. Although the cell walls of most living plant cells are quite freely permeable to water and solutes, the cytoplasmic layer that lines the cell wall is more permeable to some substances than to others. This property of differential permeability appears to reside principally in the layer of cytoplasm adjacent to the cell wall (plasma membrane, or plasmalemma) and in the layer in contact with the vacuole (vacuolar membrane, or tonoplast). This cytoplasmic system of membranes is usually relatively permeable to water, to dissolved gases, and to certain dissolved organic components. It is often much less permeable to sugars and mineral salts. The permeability of the cytoplasmic membranes is quite variable, however, and under certain metabolic conditions solutes that ordinarily penetrate through these membranes slowly or not at all may pass into or out of cells rapidly. See CELL (BIOLOGICAL).

**Osmotic pressure and turgor pressure.** If a plant cell in a flaccid condition—one in which the cell sap exerts no pressure against the encompassing cytoplasm and cell wall—is immersed in pure water, inward osmosis of water into the cell sap occurs. Osmosis may be defined as the diffusion of solvent molecules, usually water, across a membrane that is more permeable to the solvent than to solutes dissolved in it. The driving force in osmosis as in other diffusion phenomena, is the diffusion pressure (DP) of the diffusing molecules, in this case the water molecules. Inward osmosis of water takes place under these conditions because the diffusion pressure of the water in the cell sap is less than that of the surrounding pure water by the amount of its osmotic pressure (OP). If the osmotic pressure of the water in the cell sap is 15 atmospheres (atm), the diffusion pressure of the water in the cell sap is 15 atm less than that of pure water at the same temperature and under the same pressure. The lesser diffusion pressure of the water in the cell sap results from the presence of solutes. The gain in water by the cell results in the exertion of a turgor pressure (TP) against the protoplasm which in turn is transmitted to the cell wall. This pressure also prevails throughout the mass of solu-

tion within the cell. If the cell wall is elastic some expansion in the volume of the cell occurs, although in many kinds of cells this is relatively small.

Because of the solutes invariably present, the cell sap possesses an osmotic pressure. The osmotic pressures of most plant cell saps lie within a 5- to 40-atm range of magnitudes, although values as high as 200 atm have been found in some halophytes (plants that can tolerate high-solute media). The osmotic pressures of the cells of a given plant tissue vary considerably with environmental conditions and intrinsic metabolic activities. More or less regular daily or seasonal variations occur in the magnitude of cell-sap osmotic pressures in the cells of many tissues. It is the osmotic pressure of the cell sap coupled with the differential permeability of the cytoplasmic membranes and the relative inelasticity of the cell walls which permits the development of the more or less turgid condition characteristic of most plant cells.

With continued osmosis of water into the cell, its turgor pressure gradually increases until it is equal to the final osmotic pressure of the cell sap. Subjection of the water in the cell sap to pressure increases its diffusion pressure by the amount of the imposed pressure. In the example given above, disregarding the usually small amount of sap dilution as a result of cell expansion, the diffusion pressure of the water in the cell sap is reduced 15 atm because of the presence of solutes (the osmotic pressure is the index of this lowering of diffusion pressure) and raised 15 atm as a result of turgor pressure when maximum turgor is reached. Hence when dynamic equilibrium is attained the diffusion pressure of the water in the cell sap is equal to that of the surrounding water, a condition which must necessarily obtain if an equilibrium is to be achieved.

If the same cell in a flaccid condition is immersed in a solution with an osmotic pressure of 6 atm, inward osmosis occurs, but it does not continue as long as when the cell is immersed in pure water. Disregarding sap dilution, a dynamic equilibrium will be attained under these circumstances when the turgor pressure of the cell sap has reached 9 atm because at this point the diffusion pressures of the water in the cell sap and in the surrounding solution will be equal. Since the diffusion pressure of the water in the cell sap was originally diminished 15 atm because of the presence of solutes and then raised 9 atm because of turgor pressure, the net reduction in diffusion pressure, therefore, is 6 atm, which is the same as the reduction of diffusion pressure in the surrounding solution. At dynamic equilibrium the number of water molecules entering the cell and the number leaving it per unit of time will be equal.

**Diffusion pressure deficit.** As the examples given indicate, the effective physical quantity controlling the direction of osmotic movement of water from cell to cell in plants or between a cell and an external solution is the diffusion pressure deficit

(DPD) of the water. This quantity is equal to the osmotic pressure of the water less the turgor pressure to which it is subjected. Under certain circumstances the turgor pressure may be negative in value, that is, the water may be under tension. Examples of the occurrence of water under tension in plants will be discussed later in this article. When water is under tension its diffusion pressure deficit is equal to its osmotic pressure plus the tension to which it is subjected. In an unconfined solution the diffusion pressure deficit is equal to its osmotic pressure, since there is no turgor pressure. In a fully turgid plant cell the diffusion pressure deficit is zero, and the turgor pressure is equal to the osmotic pressure; in a fully flaccid cell the turgor pressure is zero, and the diffusion pressure is equal to the osmotic pressure.

The interrelationships among the principal osmotic quantities of a plant cell can be expressed in the simple equation

$$\text{DPD} = \text{OP} - \text{TP}$$

These relationships are also illustrated diagrammatically in Fig. 1, in which allowance has also been made for the effect of volume changes which are characteristic of some kinds of cells with shifts in turgor pressure. The conditions which would prevail in the cell if the cell sap passed into a state of tension (negative pressure) are indicated by the dotted extensions of the curves to the left.

Cell-to-cell movement of water in plants always occurs from the cell of lesser to the cell of greater diffusion pressure deficit. Such movement of water in plant tissues apparently often occurs over considerable distances along diffusion pressure deficit gradients in which the diffusion pressure deficit of

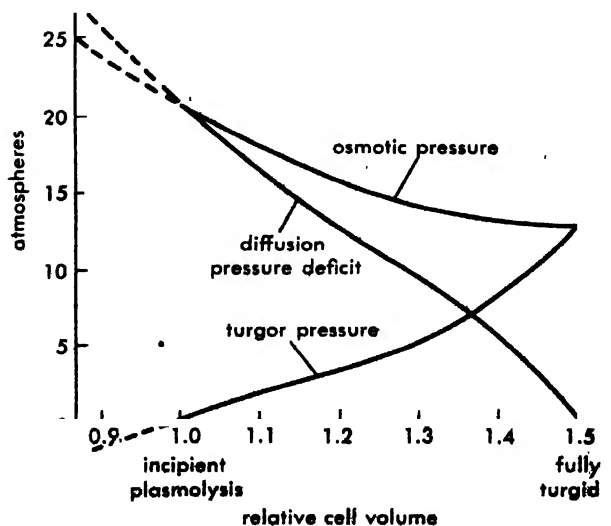


Fig. 1. Interrelationships among osmotic pressures, turgor pressures, diffusion pressure deficits, and volumes of a plant cell. (From K. Höfler, *Ber. Deut. Botan. Ges.*, 38:288-298, 1920)



each cell in a series is greater than that of the preceding one.

**Plasmolysis.** If a turgid or partially turgid plant cell is immersed in a solution with a greater osmotic pressure than the cell sap, a gradual shrinkage in the volume of the cell ensues; the shrinkage may be small or large depending upon the kind of cell and its initial degree of turgidity. When the lower limit of cell wall elasticity is reached, the protoplasmic layer begins to recede from the inner surface of the cell wall as a result of the continued loss of water from the cell sap. Retreat of the protoplasm from the cell often continues until it has shrunk towards the center of the cell, the space between the protoplasm and the cell wall becoming occupied by the bathing solution. This phenomenon is called plasmolysis. If a cell is immersed in a solution with an osmotic pressure which just slightly exceeds that of the cell sap, withdrawal of the protoplasm from the cell wall should be just barely initiated. The stage of plasmolysis shown in Fig. 1 is called incipient plasmolysis, and it is the basis for one of the commonly used methods of measuring the osmotic pressure of plant cells.

**Imbibition.** In some kinds of plant cells movement of water occurs principally by the process of imbibition rather than osmosis. The swelling of dry seeds when immersed in water is a familiar example of this process. Imbibition, like osmosis, is basically a diffusion process and occurs because of the greater diffusion pressure deficit of the water in the imbibant as compared with the diffusion pressure deficit of the water in some contiguous part of the system. An equilibrium is reached only when the diffusion pressure deficit of the water in the two parts of the system has attained the same value. The diffusion pressure deficit of the water in a dry seed is extremely high, being equal in value to its so-called "imbibition pressure"; that of pure water is zero, hence movement of water into the seed occurs. Even if the seeds are immersed in a solution of considerable osmotic pressure, which in an unconfined solution is an index of its diffusion pressure deficit, imbibition occurs. However, if the osmotic pressure of the solution is high enough (of the order of 1000 atm) the seed will not gain water from the solution, and may even lose a little to the solution. In other words, if the diffusion pressure deficit of the solution is high enough, imbibition does not occur.

However, a difference in diffusion pressure deficits between the liquid in an imbibant and in the surrounding or adjacent medium is not the only condition which must be fulfilled if imbibition is to occur. Seeds swell readily when immersed in water, but not when immersed in ether or other organic liquids. Contrariwise, rubber does not imbibe water, but does imbibe measurable quantities of ether and other organic liquids. Certain specific attractive forces between the molecules of the imbibant and the imbibed liquid are therefore also a requisite for the occurrence of imbibition.

In an imbibitional system the imbibition pressure (IP) of the imbibant is the analogue of the

osmotic pressure in an osmotic system; hence such a system

$$DPD = IP - TP$$

The imbibition pressure may be regarded as the index of the reduction in diffusion pressure in an imbibant insofar as this results from attractions between the molecules of the imbibant and water molecules. For an unconfined imbibant which is immersed in water the diffusion pressure deficit initially equals the imbibition pressure since there is no turgor pressure factor. The more nearly saturated such an imbibant becomes, the smaller is the imbibition pressure and hence also its diffusion pressure deficit. A fully saturated imbibant has zero imbibition pressure and a zero diffusion pressure deficit.

**The stomatal mechanism.** Various gases diffuse into and out of physiologically active plants. The gases of greatest physiological significance are carbon dioxide, which diffuses into the plant during photosynthesis and is lost from the plant in respiration; oxygen, which diffuses in during respiration and is lost during photosynthesis; and water vapor, which is lost in the process of transpiration. The great bulk of the gaseous exchanges between a plant and its environment occurs through tiny pores in the epidermis called stomates (see EPIDERMIS OF PLANT). Although stomates occur on many aerial parts of plants, they are most characteristic of, and occur in greatest abundance in, leaves. See LEAF (BOTANY).

Each stomate or stoma (plural, stomates or stomata) consists of a minute elliptical pore surrounded by two distinctively shaped epidermal cells called guard cells. Stomates are sometimes open and sometimes closed; when closed all gaseous exchanges between a plant and its environment are greatly retarded. The size of a fully open stomate differs greatly from one species of plant to another. Among the largest known are those of the wandering Jew (*Zebrina pendula*) whose axial dimension averages 31 by 12 microns ( $\mu$ ). In most species the stomates are much smaller, but all of them afford portals of egress or ingress which are enormous relative to the size of the gas molecules that diffuse through them. The number of stomates per square centimeter of leaf surface ranges from a few thousand in some species to over a hundred thousand in others. In many species of plants stomates are present in both the upper and lower epidermises, usually being more abundant in the lower. In many species, especially of woody plants, they are present only in the lower epidermis. In floating leaved aquatic species stomates occur only in the upper epidermis.

Rates of transpiration (loss of water vapor) from leaves of the expanded type often are 50

from leaf  
despite t  
x noon

significant is the fact that the rate of diffusion of carbon dioxide, essential in photosynthesis, into the leaves through the stomates is much greater than through the equivalent area of an efficient carbon dioxide absorbing surface.

Although some mass flow of gases undoubtedly occurs through the stomates under certain conditions, most movement of gases into or out of a leaf takes place by diffusion through the stomates. Diffusion is the physical process whereby molecules of gas move from a region of their greater diffusion pressure to the region of their lesser diffusion pressure as a result of their own kinetic activity. Molecules of liquids and solids (to a limited extent), molecules and ions of solutes, and colloidal particles also diffuse whenever the appropriate circumstances prevail. As previously pointed out, osmosis and imbibition are basically diffusion processes involving the movement of molecules in the liquid state.

Diffusion of gases through small pores follows certain principles which account for the high diffusive capacity of the stomates. In the diffusion of a gas through a small pore, an overwhelming proportion of the molecules escape over the rim of the pore relative to those escaping through its center. Hence, diffusion rates through small apertures vary as their perimeter rather than as their area. The less the area of a pore, therefore, the greater its diffusive capacity relative to its area. Therefore, a gas may diffuse nearly as rapidly through a septum pierced with a number of small orifices, whose aggregate area represents only a small proportion of the septum area, as through an open surface equal in area to the septum. The high diffusive capacity of the stomates can be accounted for in terms of these principles. Since diffusion of gases through stomates is proportional to the perimeter of the pore rather than to its area, the diffusion rate through a partially open stomate is almost as great as through a fully open stomate.

In general stomates are open in the daytime and closed at night, although there are many exceptions to this statement. The mechanism whereby stomates open in the light and close in the dark seems to be principally an osmotic one although other factors are probably involved. Upon the advent of illumination, the hydrogen-ion concentration of the guard cells decreases. This favors the action of the enzyme phosphorylase which, in the presence of phosphates, causes transformation of insoluble starch into the soluble compound glucose-1-phosphate. The resulting increase in solute concentration of the guard cells causes an increase in their osmotic pressure and hence also in their diffusion pressure deficit. Osmotic movement of water takes place from contiguous epidermal cells, in which there is little daily variation in osmotic pressure, into the guard cells. The resulting increase in the turgor pressure of the guard cells causes them to open. With the advent of darkness or of a relatively low light intensity, the reverse train of processes is apparently set in operation, leading ultimately to stomatal closure.

Light of low intensity is, generally speaking, less effective than stronger illumination in inducing stomatal opening. Hence stomates often do not open as wide on cloudy as on clear days, and often do not remain open for as much of the daylight period. A deficiency of water within the plant also induces partial to complete closure of the stomates. During periods of drought, therefore, stomates remain shut continuously or, at most, are open for only short periods each day, regardless of the light intensity to which the plant is exposed. Opening of the stomates does not occur in most species at temperatures approaching freezing. Hence in cold or even cool weather stomates often remain closed even when other environmental conditions are favorable to their opening. Nocturnal opening occurs at times in some species, but the conditions which induce this pattern of stomatal reaction are not clearly understood.

### THE PROCESS OF TRANSPIRATION

The term transpiration is used to designate the process whereby water vapor is lost from plants. Although basically an evaporation process, transpiration is complicated by other physical and physiological conditions prevailing in the plant. Whereas loss of water vapor can occur from any part of the plant which is exposed to the atmosphere, the great bulk of all transpiration occurs from the leaves. There are two kinds of foliar transpiration: (1) stomatal transpiration, in which water vapor loss occurs through the stomates, and (2) cuticular transpiration, which occurs directly from the outside surface of epidermal walls through the cuticle. In most species 90% or more of all foliar transpiration is of the stomatal type.

**Dynamics of stomatal transpiration.** The dynamics of stomatal transpiration is considerably more complex than that of cuticular transpiration. In the leaves of most kinds of plants the mesophyll cells do not fit together tightly and the intercellular spaces between them are occupied by air. A veritable labyrinth of air-filled spaces is thus present within a leaf, bounded by the water-saturated walls of the mesophyll cells. Water evaporates readily from these wet cell walls into the intercellular spaces. If the stomates are closed, the only effect of such evaporation is to saturate the intercellular spaces with water vapor. If the stomates are open, however, diffusion of water vapor usually occurs through them into the surrounding atmosphere. Such diffusion always occurs unless the atmosphere has a vapor pressure equal to or greater than that within the intercellular spaces, a situation which seldom prevails during the daylight hours of clear days. The two physical processes of evaporation and diffusion of water vapor are both integral steps in stomatal transpiration. Physiological control of this component of transpiration is exerted through the opening and closing of the stomates, previously described.

**Effects of environment on transpiration.** Light is one of the major factors influencing the rate of transpiration because of its controlling effect on

the opening and closing of stomates. Since stomatal transpiration is largely restricted to the daylight hours, daytime rates of transpiration are usually many times greater than night rates, which largely or entirely represent cuticular transpiration. Since leaves in direct sunlight usually have temperatures from one to several degrees higher than that of the surrounding atmosphere, light also has a secondary accelerating effect on transpiration through its influence on leaf temperatures. Increase in leaf temperature results in an increase in the diffusion pressure of the water vapor molecules within the leaf.

The rate of diffusion of water vapor through open stomates depends upon the steepness of the vapor pressure gradient between the intercellular spaces and the outside atmosphere. When the vapor pressure in that part of the intercellular spaces just below the stomatal pores is high relative to that of the atmosphere, diffusion of water vapor out of the leaf occurs rapidly; when it is low, water vapor diffusion occurs much more slowly.

Temperature has a marked effect upon rates of transpiration principally because of its differential effect upon the vapor pressure of the intercellular spaces and atmosphere. Although leaf temperatures do not exactly parallel atmospheric temperatures, increase in atmospheric temperature in general results in a rise in leaf temperature and vice versa. On a warm, clear day such as would be typified by many summer days in temperate latitudes, and with an adequate soil water supply, increase in temperature results in an increase in the vapor pressure of the intercellular spaces. Such a rise in vapor pressure occurs because the vapor pressure corresponding to a saturated condition of an atmosphere increases with rise in temperature, and the extensive evaporating surfaces of the cell walls bounding the intercellular spaces make it possible for the intercellular spaces to be maintained in an approximately saturated condition most of the time. An increase in temperature ordinarily has little or no effect on the vapor pressure of the atmosphere and this is especially true of warm, bright days on which transpiration rates are the highest. Hence, as the temperature rises, the vapor pressure of the intercellular spaces increases relative to that of the external atmosphere, the vapor pressure gradient through the stomates is steepened, and the rate of outward diffusion of water vapor increases.

Wind velocity is another factor which influences the rate of transpiration. Generally speaking, a gentle breeze is relatively much more effective in increasing transpiration rates than are winds of greater velocity. In quiet air localized zones of relatively high atmospheric vapor pressure may build up in the vicinity of transpiring leaves. Such zones retard transpiration unless there is sufficient air movement to prevent the accumulation of water vapor molecules. The bending, twisting, and fluttering of leaf blades and the swaying of stalks and branches in a wind also contribute to increasing the rate of transpiration.

Soil water conditions exert a major influence on the rate of transpiration. Whenever soil conditions

are such that the rate of absorption of water is retarded there is a corresponding diminution in the rate of transpiration.

**Daily periodicity of transpiration.** The rate of every major plant process, including transpiration, is measurably and often markedly influenced by the environmental conditions to which the plant is exposed. Many of the environmental factors exhibit more or less regular daily periodicities, which vary somewhat, of course, with the prevailing climatic conditions. This is especially true of the factors of light and temperature. Many plant processes, including transpiration, therefore exhibit daily periodicities in rate that are correlated with the daily periodicities of one or more environmental factors.

The daily periodicity of transpiration in alfalfa as exhibited on a clear, warm day with adequate soil water available is illustrated in Fig. 2. A similar daily periodicity of transpiration is exhibited under comparable environmental conditions by many other species. During the hours of darkness the transpiration rate is relatively low, and in most species water vapor loss during this period may be regarded as entirely cuticular or nearly so. The transpiration rate shows a steady rise during the morning hours culminating in a peak rate which is attained in the early hours of the afternoon. The increase in transpiration rate during the forepart of the day results from gradual opening of the stomates beginning with the advent of light, followed by a steady increase in the steepness of the vapor pressure gradient through the stomates, which occurs as a result of increasing atmospheric temperature during the morning and earlier afternoon hours.

In most plants, if transpiration is occurring rapidly, the rate of absorption of water does not keep pace with the rate at which water vapor is lost from the leaves. In other words the plant is gradually being depleted of water during the daylight hours. In time the resulting decrease in the water content of the leaf cells results in a reduced vapor pressure within the intercellular spaces, and a diminution in the rate of transpiration begins. Stomates also start to close as a result of the diminished leaf water content, and their closure is accelerated during the latter part of the afternoon by the waning light intensity. By nightfall complete closure of virtually all stomates has taken place, and water vapor loss during the hours of darkness is again restricted

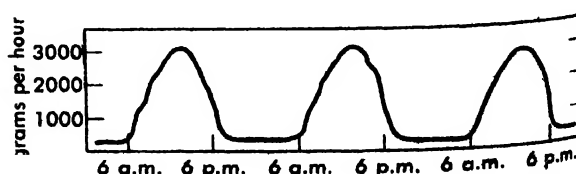


Fig. 2. Daily periodicity of transpiration of alfalfa on three successive clear, warm days with adequate soil water. Rate of transpiration expressed as grams per hour per 6 ft square plot of alfalfa. (From M. D. Thomas and G. R. Hill, *Plant Physiol.*, 12:2 1937)

largely or entirely to the relatively low rate of cuticular transpiration. It is noteworthy that the peak rate of transpiration occurs during the early afternoon hours, correlating more closely with the daily temperature periodicity than with the daily periodicity of light intensity.

Under environmental conditions differing considerably from those postulated in the preceding discussion, patterns of transpiration periodicity may show a considerable variance from the one described. On cloudy days, for example, stomates generally open less completely than on clear days, and a curve for daily transpiration periodicity presents a greatly flattened appearance as compared with the one shown in Fig. 2. A cool temperature, even in a range somewhat above freezing, greatly diminishes and may even result in cessation of stomatal transpiration, resulting in a pronounced modification in the daily march of transpiration periodicity. A deficient soil water supply is probably the most common cause of departures from the pattern of transpiration described above. A reduction in soil water content below the field capacity (optimum water availability) results not only in a general flattening of the transpiration periodicity curve, but frequently also in appearance of the peak of the curve somewhat earlier in the day. Since even in temperate zone regions, drought periods of greater or less severity are of common occurrence during the summer months, and in many habitats are the rule rather than the exception, transpiration periodicity curves of this flattened, and skewed peak type are undoubtedly of frequent occurrence.

**Magnitude of transpiration.** Transpiration of broad leaved species of plants growing under temperate zone conditions may range up to about 5 grams per square decimeter of leaf area per hour. Sufficient quantities of water are often lost in transpiration by vegetation-covered areas of the earth's surface to have important effects not only on soil water relations, but also on meteorological conditions. The quantities of water lost per acre by crops, grasslands, or forest are therefore a matter of basic interest. An acre of corn (maize), for example, transpires water equivalent to 15 in. of rainfall during a usual growing season. Transpiration of deciduous, largely oak, forest in the southern Appalachian mountains has been estimated as equivalent to 17-22 in. of rainfall per year. Marked variations occur in such values from year to year, however, depending upon prevailing climatic conditions.

**Significance of transpiration.** Viewpoints regarding the significance of transpiration have ranged between the two extremes of considering it a process that is (1) an unavoidable evil, or (2) a physiological necessity. Neither of these extreme views appears to be tenable. Some of the incidental effects of transpiration appear to be advantageous to the plant, but none of them is indispensable for its survival or even for its adequate physiological operation. Likewise, while some of the incidental effects of transpiration appear to be detrimental to

the plant, none of them is so in such a critical fashion that survival of plants, considered in the aggregate, is endangered.

Transpiration is a necessary consequence of the relation of water to the anatomy of the plant, and especially to the anatomy of the leaves. Terrestrial green plants are dependent upon atmospheric carbon dioxide for their survival. In terrestrial vascular plants the principal carbon dioxide absorbing surfaces are the moist mesophyll cell walls which bound the intercellular spaces in leaves. Ingress of carbon dioxide into these spaces occurs mostly by diffusion through open stomates. When the stomates are open, outward diffusion of water vapor unavoidably occurs, and such stomatal transpiration accounts for most of the water vapor loss from plants. Although transpiration is thus, in effect, an incidental phenomenon, it often has marked indirect effects on other physiological processes occurring in the plant because of its effects on the internal water relations of the plant.

#### TRANSLOCATION OF WATER

In terrestrial rooted plants practically all of the water which enters a plant is absorbed from the soil by the roots. The water thus absorbed is translocated to all parts of the plant. In the tallest trees (specimens of the coast redwood, *Sequoia sempervirens*) the distance from the tips of the deepest roots to the tips of the topmost branches is nearly 400 ft, and water must be elevated for this distance through such trees. Although few plants are as tall as such redwoods, the same mechanisms of water movement are believed to operate in all vascular species. The mechanism of the "ascent of sap" (all translocated water contains at least traces of solutes) in plants and especially tall trees, was one of the first processes to excite the interest of plant physiologists.

**Pathway of water movement.** The upward movement of water in plants occurs in the xylem, which, in the larger roots, trunks, and branches of trees and shrubs, is identical with the wood (see XYLEM). In the trunks or larger branches of most kinds of trees, however, sap movement is restricted to a few of the outermost annual layers of wood. This explains why hollow trees, in which the central core of older wood has disintegrated, can remain alive for many years. The xylem of any plant is a unit and continuous system throughout the plant. Small strands of this tissue extend almost to the tip of every root. Other strands, the larger of which constitute important parts of the veins, ramify to all parts of each leaf. In angiosperms most translocation of water occurs through the xylem vessels, which are nonliving, elongated, tubelike structures. The vessels are formed by the end-to-end coalescence of many much smaller cells, death of these cells ensuing at about the same time that coalescence occurs. In trees the diameters of such vessels range from about 20 to 400  $\mu$ , and they may extend for many feet with no more interruption than an occasional incomplete cross-wall. In gymnosperms no vessels are present and movement of

water occurs solely through spindle-shaped xylem cells called tracheids. Vertically contiguous tracheids always overlap along their tapering portions, resulting in a densely packed type of woody tissue. Individual tracheids may be as much as 5 mm in length. Like the vessels, they are nonliving while functional in the translocation of water. Small, more or less rounded, thin areas occur in the walls of vessels and tracheids that are contiguous with the walls of other tracheids, vessels, or cells. Structurally three main types of such pits are recognized, but all of them appear to facilitate the passage of water from one xylem element to another.

**Root pressure.** The exudation of xylem sap from the stump of a cut off herbaceous plant is a commonly observed phenomenon. Sap exudation ("bleeding") from the cut ends of stems or from incisions into the wood also occurs in certain woody plants, such as birch, currant, and grape, especially in the spring. A single vigorous grapevine often loses a liter or more of sap per day through the cut ends of stems after spring pruning. This exudation of sap from the xylem tissue results from a pressure originating in the roots, called root pressure. A related phenomenon is that of guttation. This term refers to the exudation of drops of water from the tips or margins of leaves and occurs in many species of herbaceous plants as well as in some woody species. Like sap exudation from cut stems, this phenomenon is observed most frequently in the spring, and especially during early morning hours. The water exuded in guttation is not pure, but contains traces of sugar and other solutes. Guttation occurs from special structures called hydathodes which are similar in structure to, but larger, than stomates. In most species water loss by guttation is negligible in comparison with the water lost as vapor in transpiration. Like xylem sap exudation, guttation results from root pressure.

Root pressure is generally considered to be one of the mechanisms of upward transport of water in plants. While it is undoubtedly true that root pressure does account for some upward movement of water in certain species of plants at some seasons, various considerations indicate that it can be only a secondary mechanism of water transport. Among these are (1) there are many species in which the phenomenon of root pressure has not been observed, (2) the magnitude of measured root pressures seldom exceeds 2 atm, which could not activate a rise of water for more than about 60 ft, and many trees are much taller than this, (3) known rates of xylem-sap flow under the influence of root pressure are usually inadequate to compensate for known rates of transpiration, (4) root pressures are usually operative in woody plants only during the early spring; during the summer months when transpiration rates, and hence rates of xylem-sap transport, are greatest, root pressures are negligible or nonexistent.

**Cohesion of water and ascent of sap.** Although invariably in motion, as a result of their kinetic energy, water molecules are also strongly attracted to

each other. In masses of liquid water the existence of such intermolecular attractions is not obvious, but when water is confined in long tubes of small diameter the reality of the mutual attractions among water molecules can be demonstrated. If the water at the top of such a tube be subjected to a pull the resulting stress will, because of the mutual attraction (cohesion) among water molecules, be transmitted all the way down the column of water. Furthermore, because of the attraction between the water molecules and the wall of the tube (adhesion), subjecting the water column to a stress does not result in pulling it away from the wall.

The facts just recited have been made the basis of a widely entertained theory of the mechanism of water transport in plants, first clearly enunciated by H. H. Dixon in 1914. According to this theory, upward translocation of water is engendered by the development of diffusion pressure deficits in the cells of apical organs of plants. Such diffusion pressure deficits develop most commonly in the mesophyll cells of leaves, hence this concept of the mechanism of water translocation is usually associated with the process of transpiration.

Evaporation of water from the walls of the mesophyll cells abutting on the intercellular spaces results in an increase in the diffusion pressure deficit of these cells. Consequent cell-to-cell movements of water cause an increase in the diffusion pressure deficit even of those mesophyll cells which are not directly exposed to the intercellular spaces. The resulting increase in diffusion pressure deficit of those cells directly in contact with the xylem elements in the veinlets of the leaf induces movement of water from the vessels or tracheids into these adjacent cells. Since, whenever transpiration is occurring at appreciable rates, water does not enter the lower ends of the xylem conduits in the roots as rapidly as it passes out of the vessels or tracheids into adjacent cells at the upper ends of the water-conductive system, the water in the xylem ducts is stretched into taut threads, that is, it passes into a state of tension. Each column of water behaves like a tiny stretched wire. The tension is transmitted along the entire length of the water columns to their terminations just back of the root tips. Since the tension sustained by the water in the xylem ducts is in effect a diffusion pressure deficit (the osmotic factor in the diffusion pressure deficit of xylem sap is usually small relative to the tension factor), movement of water is induced from adjacent root cells into xylem elements in the absorbing regions of roots.

The tension engendered in the water columns can be sustained by them because of the cohesion between the water molecules, acting in conjunction with the adhesion of the boundary layers of water molecules to the walls of the xylem ducts. The existence of water under tension in vessels has been verified in a number of species of plants by direct microscopic examination. There is some evidence that, under conditions of marked internal water deficiency, the tensions generated in the water col-



umns are proliferated into the mesophyll cells of leaves and cells in the absorbing regions of roots. Conservative calculations indicate that a cohesion value of 30–50 atm would be adequate to permit translocation of water to the top of the tallest known trees. However, under conditions of internal water deficiency, tensions considerably in excess of 50 atm are probably engendered in the water columns of many plants, especially woody species.

#### ABSORPTION OF WATER

This process will be discussed only from the standpoint of terrestrial, rooted plants. Consideration of the absorption of water by plants necessitates an understanding of the physical status of the water in soils as it exists under various conditions.

**Soil water conditions.** Even in the tightest of soils the particles never fit together perfectly and a certain amount of space exists among them (see SOIL). This pore space of a soil ranges from about 30% of the soil volume in sandy soils to about 50% of the soil volume in heavy clay soils. In desiccated soils the pore space is occupied entirely by air, in saturated soils it is occupied entirely by water, but in moist, well-drained soils it is usually occupied partly by air and partly by water. In a soil in which a water table is located not too far below the surface, considerable quantities of water may rise into its upper layers by capillarity and become available to plants. In arid regions, however, there ordinarily is no water table. Even in many humid regions the water table is continuously or intermittently too far below the soil surface to be an appreciable source of water for most plants. In all soils lacking a water table, or in which the water table is at a considerable depth, the only water available to plants is that which comes as natural precipitation or which is provided by artificial irrigation. If water falls on or is applied to a dry soil which is homogeneous to a considerable depth, it will become rapidly distributed to a depth which will depend on the quantity of water supplied per unit area, and on the specific properties of that soil. After several days, further deepening of the moist layer of soil extending downward from the surface virtually ceases, because within such a time interval capillary movement of water in a downward direction has become extremely slow or nonexistent. The boundary line between the moist layer of soil above and the drier zone below will be a distinct one. In this condition of field equilibrium the water content of the upper moist soil layer is, in homogeneous soils, essentially uniform throughout.

The water content of a soil in this equilibrium condition is called the field capacity. Field capacities range from about 5% in coarse sandy soils to about 45% of the dry weight in clay soils. The moisture equivalent of a soil, often measured in the laboratory, is usually very close in value to the field capacity of the same soil. It is defined as the water content of a soil which is retained against a force 1000 times gravity as measured in a centrifuge. A soil at its field capacity is relatively moist, but is

also well aerated. Soil water contents at or near the field capacity are the most favorable for growth of most kinds of plants.

A considerable proportion of the water in any soil is unavailable in the growth of plants. The permanent wilting percentage is the generally accepted index of this fraction of the soil water. This quantity is measured by allowing a plant to develop with its roots in soil enclosed in a waterproof pot until the plant passes into a state of permanent wilting. The water content of the soil when the plant just passes into this condition is the permanent wilting percentage. The range of permanent wilting percentages is from 2–3% of the dry weight in coarse sandy soils to about 20% in heavy clay loams. About the same value is obtained for the permanent wilting percentage of a given soil, regardless of the kind of test plant used.

The diffusion pressure deficit of the soil water has two major components. One is the osmotic pressure of the soil solution, which in most kinds of soils is only a fraction of an atmosphere. The other is the attractive forces between the soil particles and water molecules which may attain a very considerable magnitude, especially in dry soils. In moist soils, those at the field capacity or higher soil water content, the former of these two components is principally responsible for the soil water diffusion pressure deficit; in drier soils the latter of these two components is almost solely responsible for the diffusion pressure deficit of the soil water. In the majority of soils the diffusion pressure deficit of the soil water is close to zero at saturation, less than 1 atm at field capacity, and in the vicinity of 15 atm at the permanent wilting percentage. With further reduction in the water content of a soil below its permanent wilting percentage, its diffusion pressure deficit increases rapidly and at an accelerating rate. Almost no water can be absorbed by plants at such high soil water diffusion pressure deficits; it is for this reason that the permanent wilting percentage is the index of the soil water which is unavailable in plant growth (Fig. 3).

**Relation of root growth to water absorption.** The successively smaller branches of the root system of any plant terminate ultimately in the root tips, of which there may be thousands and often millions on a single plant. As generally employed, the term root tip refers to the region extending back from the apex of the root for a distance of at least several centimeters. The terminal zone of a root tip is the root cap. Just back of this are the regions in which cell division and cell elongation occur and in which all growth in length of roots takes place. See ROOT (BOTANY). Just back of these regions, in the majority of species, is the zone of root hairs. Each root hair is a projection from the epidermal cell of which it is an integral part. A single root tip may bear thousands of root hairs, ranging in length from a few millimeters up to about a centimeter. In most species the root hairs are short-lived structures, but new ones are con-



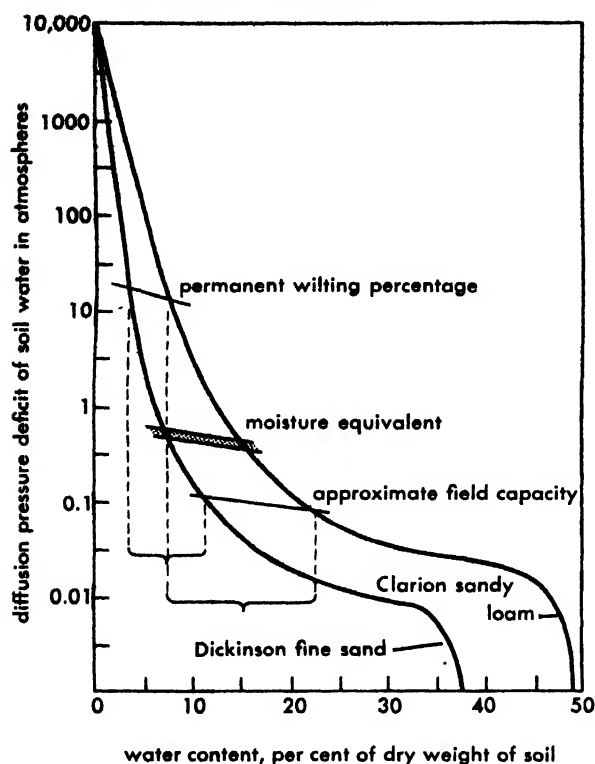


Fig. 3. Relation between diffusion pressure deficit of soil water and soil water content of two soils over entire range of soil water contents. Horizontal brackets show the range for each soil over which water is readily available to plants. The vertical scale is logarithmic. (From M. B. Russell, *Am. Soil Sci. Soc. Proc.*, 4:51-54, 1939)

stantly developing just back of the growing region of the root as it elongates. Most absorption of water occurs in the root tip regions, and especially in the root hair zone. Older portions of most roots become covered with cutinized or suberized layers through which only very limited quantities of water can pass.

Whenever the diffusion pressure deficit of the water in the root hairs and other peripheral cells of a root tip exceeds that of the soil water, movement of water takes place into the root cells. If the soil water content exceeds the field capacity, water may move by capillarity toward the region of absorption from portions of the soil not immediately contiguous with the root tips, and the supply of readily absorbable water is maintained in this way. Elongation of the roots, although slower in most species in relatively wet soils, also helps maintain contact between the root tips and untapped portions of the soil water. Many plants, much of the time, grow in soils with a water content in the range between the field capacity and the permanent wilting percentage. In this range of soil water contents, capillary movement of water through the soil is extremely slow or nonexistent, and an adequate supply of water cannot be maintained to rapidly absorbing root tips by this means. In such soils maintenance of contact between the root tips and

available soil water is assured only by continued elongation of the roots through the soil. Mature root systems of many plants terminate in millions of root tips, each of which may be visualized as slowly advancing through the soil, absorbing water from around or between the soil particles with which it comes in contact. The aggregate increase in the length of the root system of a rye plant averages 3.1 mi/day. Calculations indicate that the daily root elongation of this plant is adequate to permit absorption of a sufficient quantity of water from soils at the field capacity to compensate for daily transpirational water loss.

**Mechanisms of absorption of water.** As previously indicated, the tension generated in the water columns of a plant, most commonly as an indirect result of transpiration, is transmitted to the ultimate terminations of the xylem ducts in the root tips. As soon as the tension in the water columns exceeds the diffusion pressure deficit of contiguous cells in the root tip, water moves from those cells into the xylem. This activates further cell-to-cell movement of water in a lateral direction across the root and presumably in the establishment of a gradient of diffusion pressure deficits, increasing progressively in magnitude from the epidermal cells, including the root hairs, to the root xylem. Whenever the diffusion pressure deficit of the peripheral cells of the root exceeds that of the soil water, movement of water from the soil into the root cells occurs. There is some evidence that under conditions of marked internal water stress, the tension generated in the xylem ducts will be propagated across the root to the peripheral cells. If this occurs, greater diffusion pressure deficits could develop in peripheral root cells than would otherwise be possible. This absorption mechanism would operate in fundamentally the same way whether or not the water in the root cells passes into a state of tension. The process just described, often called passive absorption, accounts for most of the absorption of water by terrestrial plants.

The phenomenon of root pressure, previously described as the basis for xylem sap exudation from cuts or wounds and guttation, represents another mechanism of the absorption of water. This mechanism is localized in the roots and is often called active absorption. Water absorption of this type only occurs when the rate of transpiration is low and the soil is relatively moist. Although xylem sap is relatively dilute, its osmotic pressure is usually greater than the diffusion pressure deficit of the soil water when the soil is relatively moist. A gradient of diffusion pressure deficits can thus be established across the cortex and other tissues of the root along which the water moves laterally from the soil to the xylem. There is evidence, however, that a respiration mechanism as well as an osmotic mechanism may be involved in the correlated phenomena of active absorption, root pressure, and guttation.

**Effects of environment on absorption.** Any factor which influences the rate of transpiration also

influences the rate of absorption of water by plants and vice versa. Climatic conditions may therefore indirectly affect rates of water absorption, and soil conditions indirectly affect transpiration. Low soil temperatures, even in a range considerably above freezing, retard the rate of absorption of water by many species. The rate of water absorption by sunflower plants, for example, decreases rapidly as the soil temperature drops below 55°F.

Within limits, the greater the supply of available soil water, the greater the possible rate of water absorption. High soil water contents, especially those approaching saturation, result in decreased water absorption rates in many species because of the accompanying deficient soil aeration. In the atmosphere of such soils the oxygen concentration is lower and the carbon dioxide concentration is higher than in the atmosphere proper. In general, the deficiency of oxygen in such soils appears to be a more significant factor in causing retarded rates of water absorption than the excess of carbon dioxide. This retarding effect on the rate of water absorption is correlated with a retarding effect on the rate of root respiration (see PLANT RESPIRATION).

Likewise, if the soil solution attains any considerable concentration of solutes, water absorption by the roots is retarded. In most soils the concentration of the soil solution is so low that it is a negligible factor in affecting rates of water absorption. In saline or alkali soils, however, the concentration of the soil solution may become equivalent to many atmospheres, and only a few species of plants are able to survive when rooted in such soils.

**Wilting.** Daily variations in the water content of plants, more marked in some organs than in others, are of frequent occurrence. The familiar phenomenon of wilting, exhibited by the leaves and sometimes other organs of plants, particularly herbaceous species, is direct visual evidence of this fact. In hot, bright weather the leaves of many species of plants often wilt during the afternoon, only to regain their turgidity during the night hours, even if no additional water is provided by rainfall or irrigation. This type of wilting reaction is referred to as temporary or transient wilting and clearly results from a rate of transpiration in excess of the rate of water absorption during the daylight hours. As a result the total volume of water in the plant shrinks, although not equally in all organs or tissues. In general, diminution in water content is greatest in the leaf cells, and wilting is induced whenever the turgor pressure of the leaf cells is reduced sufficiently.

Even on days when visible wilting is not discernible, incipient wilting is of frequent occurrence. Incipient wilting corresponds to only a partial loss of turgor by the leaf cells and does not result in visible drooping, folding, or rolling of the leaves. Leaves entering into the condition of transient wilting always pass first through the stage of incipient wilting. Occurrence of this invisible first stage of

is almost universal on bright, warm days

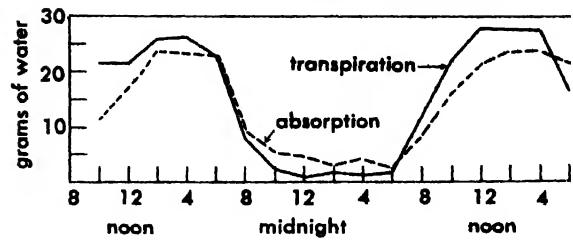


Fig. 4. Comparative daily periodicities of transpiration and absorption of water in loblolly pine (*Pinus taeda* L.). (From P. J. Kramer, *Plant and Soil Water Relationships*, McGraw-Hill, 1937)

on which environmental conditions are not severe enough to induce the more advanced stage of transient wilting.

Confirmation of the inferred cause of transient wilting has been furnished by investigations of the comparative daily periodicities of transpiration and the absorption of water (Fig. 4). As illustrated in this figure, there is a distinct lag in the rate of absorption of water as compared with the rate of transpiration during the daylight hours. During the night hours, on the contrary, the rate of water absorption is continuously greater than the rate of transpiration. Thus, during the daylight hours the tissues of the plant are being progressively depleted of water, whereas the store of water within the plant is being steadily replenished during the night hours. The lag in the rate of absorption behind the rate of transpiration during the daylight hours appears to result largely from the relatively high resistance of the living root cells to the passage of water across them.

Both incipient and transient wilting should be distinguished from the more drastic stage of permanent wilting. This stage of wilting is attained only when there is an actual deficiency of water in the soil, and a plant will not recover from permanent wilting unless the water content of the soil in which it is rooted increases. In a soil which is gradually drying out, transient wilting slowly grades over into permanent wilting. Each successive night recovery of the plant from temporary wilting takes longer and is less complete, until finally even the slightest recovery fails to take place during the night.

Although the stomates are generally closed during permanent wilting, cuticular transpiration continues. Plants in a state of permanent wilting continue to absorb water, but at a slow rate. Restoration of turgidity is not possible, however, because the rate of transpiration even from a wilted plant exceeds the rate of absorption of water from a soil at the permanent wilting percentage or lower water content. During permanent wilting, therefore, there is a slow but steady diminution in the total volume of water within the plant, and a gradual intensification of the stress in the hydrodynamic system. Tensions in the water columns of permanently wilted plants are relatively high and have been estimated to attain values of 200 atm in some trees, although this is probably an extreme figure.

As previously mentioned, the permanent wilting percentage, an important index of soil water conditions, is defined as the soil water content when a plant just enters the condition of permanent wilting. Sunflower is the most commonly used test plant in making determinations of the permanent wilting percentage of a soil. Permanent wilting of the basal pair of leaves, judged to have occurred when they fail to recover if placed in a saturated atmosphere overnight, is taken as the critical point in the measurement. The range of soil water contents between the first permanent wilting of the basal leaves of sunflower plants and the permanent wilting of all the leaves is called the wilting range. The water content of the soil at the time all the sunflower leaves have become wilted is termed the ultimate wilting point. In general the wilting range is narrower in coarse-textured soils than in fine-textured soils, and may be 10–30% of the soil water content between the field capacity and the ultimate wilting point. Although plants cannot grow while the soil in which they are rooted is in the wilting range, many kinds of plants can survive for considerable periods under such conditions. This is especially true of many shrubby species indigenous to semi-desert areas.

**Internal redistributions of water.** For convenience, the processes of transpiration, translocation of water, and absorption of water are often discussed separately, although there is a close inter-relationship among these three processes. The hydrodynamic system of a plant is essentially a unit in its operation, and changes in the status of the water in one part of a plant are bound to have effects on its status in other parts of the plant.

Whenever a plant is saturated, or nearly so, with water, differences in diffusion pressure deficits from one organ or tissue to another are minimal in value. But whenever the rate of absorption of water lags behind the rate of transpiration, an internal water deficit develops in the hydrodynamic system of the plant, which in turn favors the establishment of marked differences in diffusion pressure deficit from one part of the plant to another. Under such conditions redistribution of some of the water present from some tissues or organs of a plant to others generally occurs.

Internal movements of water from fruits to leaves and vice versa seem to be of common occurrence. Mature lemon fruits, while still attached to the tree, exhibit a daily cycle of expansion and contraction (Fig. 5). The lemon fruits begin to contract in volume early in the morning and continue to do so until late afternoon. Since transpirational loss from a lemon fruit is negligible, it is obvious that during this part of the day, corresponding to the period of high transpiration rates from the leaves, water is moving out of the fruits into other parts of the tree. Most of this movement probably occurs into the leaves. During the daylight hours the diffusion pressure deficits of the leaf cells presumably increase until they soon exceed those of the fruit cells thus initiating movement of water from fruits to leaves. During the late afternoon and night hours, the volume of the fruits gradually increases, indicating that water is now moving back into the fruits. During this period transpirational water loss from the leaves is small, leaf water contents increase, and the diffusion pressure deficit of the leaf cells diminishes. Less of the absorbed water is translocated to the leaves than during the daylight hours, and more can move into the fruits, despite the relatively low diffusion pressure deficit of the cells of the fruit. Marked daily variations take place in the diameters of lemon fruits even under environmental conditions which result in no observable wilting of the leaves.

In growing cotton bolls, however, as long as enlargement is continuing, increase in diameter continues steadily both day and night and even during periods when the leaves are severely wilted. Movement of water is obviously occurring into the growing bolls without interruption during this period. Once the bolls cease enlarging, however, reversible daily changes occur in their diameter, similar in pattern to those which take place in mature lemon fruits (Fig. 6). Similarly, it has been shown that in a tomato plant the topmost node within which growth in length occurs continues to elongate at approximately the same rate both day and night. The stem below the first node, however, shrinks measurably in length during the daytime and elongates equally at night, undoubtedly as a result of reversible changes in the turgidity of the stem cells.

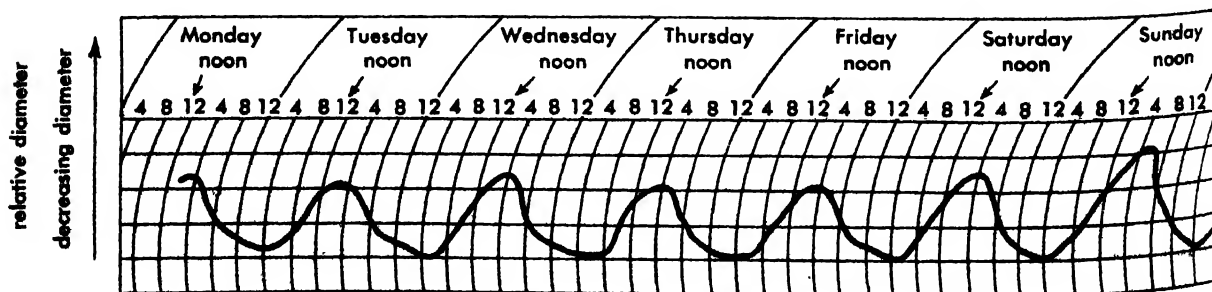


Fig. 5. Daily variations in the diameter of lemon fruits. (From E. T. Bartholomew, *Am. J. Botany*, 13:102–117, 1926)

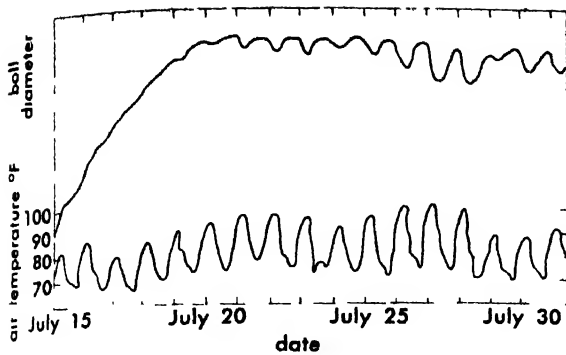


Fig 6 Daily variations in diameter of a cotton boll. During the first 5 days, the boll was still growing. (From D. B. Anderson and T. Kerr, *Plant Physiol.*, 18:261-269, 1943)

The growing cells in the terminal node of the tomato stem obviously continue to obtain water during the daylight hours while the rest of the stem is losing water, and some of the water utilized in their growth probably comes from the cells of lower nodes. In general, as the last two examples illustrate, actively meristematic regions such as growing stem and root tips and enlarging fruits, under conditions of internal water deficiency, apparently develop higher diffusion pressure deficits than other tissues (see MERISTEM, APICAL). Hence water often continues to move toward such regions even when an internal water deficiency of considerable magnitude has developed within the plant. However, under conditions of drastic internal water deficit, approaching or corresponding to a state of permanent wilting, growth of all meristems is greatly retarded or inhibited.

**Drought resistance.** The term drought refers in general to periods during which the soil contains little or no water which is available to plants. In relatively humid climates such periods are infrequent and seldom of long duration except in certain local habitats. The more arid a climate, in general the more frequent the occurrence of droughts, and the longer their duration. Most species of plants can survive short dry periods without serious injury, but a prolonged period of soil water deficiency is highly injurious or lethal to all except those species of plants with a well-developed capacity for drought resistance.

Most species which grow in semidesert regions, such as those of the southwestern United States and adjacent Mexico, or in locally dry habitats, must be drought resistant in one sense of the word or another. Annual species, which grow in many arid regions during short rainy spells, are exceptions to this statement. Such species complete their brief life cycle from seed to seed during a period when soil water is available and can survive in arid regions because they evade rather than endure drought.

Succulents, such as cacti, are found in most semidesert regions, and are also often indigenous to lo-

cally dry habitats such as sand dunes and beaches in humid climate regions. Succulents are able to survive long dry periods because of the relatively large quantities of water which accumulate, in some species in fleshy stems, and in other species in fleshy leaves, during the occasional periods when soil water is available. Many succulents can live for months on such stored water.

Those plants which are drought resistant in the truest sense are those whose cells can tolerate a marked reduction in water content for extended periods of time without injury. Many shrubby species of semidesert regions have this property. Certain structural features undoubtedly aid in the survival of such plants for long periods in arid habitats. Many xerophytes (plants that can endure periods of drought) have extensive root systems in proportion to their tops; such a structural characteristic aids in maintaining a supply of water to the aerial portions of the plant longer than would otherwise be possible (see ECOLOGY). Other drought-resistant species are characterized by having diminutive leaves; the transpiring surface of the plant may thus be small relative to the absorptive capacity of the roots. In still other species the leaves abscise (fall) with the advent of the dry season, thus greatly reducing transpiration per plant during the period of greatest internal stress in the hydrodynamic system.

Despite any structural features which may help maintain their internal water supply, shrubby plants of semiarid regions regularly undergo a gradual depletion in the store of water within them and a gradual intensification of the stress prevailing in the internal hydrodynamic system over dry periods often lasting for months. Only drought-resistant species can endure this condition, which is in essence a state of permanent wilting, for long periods of time without injury. A fundamental factor in the drought resistance of plants therefore appears to be a capacity of the cells to endure a substantial reduction in their water content without suffering injury. This capacity probably is based in part on structural features of the cells of such species and in part on the distinctive physiological properties of their protoplasm. See PLANT, MINERAL NUTRITION OF. [B.S.M.]

**Bibliography:** A. S. Crafts, H. B. Currier, and C. R. Stocking, *Water in the Physiology of Plants*, 1949; H. H. Dixon, *Transpiration and the Ascent of Sap in Plants*, 1914; P. J. Kramer, *Plant and Soil Water Relationships*, 1949; N. A. Maximov, *The Plant in Relation to Water*, 1929; B. S. Meyer and D. B. Anderson, *Plant Physiology*, 2d ed., 1952; B. T. Shaw (ed.), *Soil Physical Conditions and Plant Growth*, 1952.

## Plant anatomy

The area of plant science concerned with the internal structure of plants. It deals both with mature structures and with their origin and development.

The plant anatomist dissects the plant and studies it from different planes and at various levels of magnification. At highest magnification he examines the smallest units of plant structure, the cells; at intermediate magnification he observes the organized aggregations of these cells, the tissues; and at low magnification he determines the arrangement and interrelations of tissues in plant organs such as root, stem, leaf, and flower. At the level of the cell, anatomy overlaps plant cytology, which deals exclusively with the cell and its contents. Sometimes the name plant histology is applied to the area of plant anatomy directed toward the study of cellular details of tissues. See PLANT ORGANS; PLANT TISSUE SYSTEMS. [K.E.]

**Bibliography:** E. Boureau, *Anatomie Végétale*, 3 vols., 1954–1957; H. A. De Bary (tr. F. O. Bower and D. H. Scott), *Comparative Anatomy of the Vegetative Organs of the Phanerogams and Ferns*, 1884; A. J. Eames and L. H. MacDaniels, *An Introduction to Plant Anatomy*, 2d ed., 1947; K. Esau, *Plant Anatomy*, 1953; A. S. Foster, *Practical Plant Anatomy*, 2d ed., 1949; A. S. Foster and E. M. Gifford, Jr., *Comparative Morphology of Vascular Plants*, 1959; G. F. J. Haberlandt, *Physiological Plant Anatomy*, 1914; B. D. Jackson, *A Glossary of Botanic Terms*, 4th ed., 1953; H. Lundegårdh, *Zelle und Cytoplasma*, in K. Linsbauer, *Handbuch der Pflanzenanatomie*, vol. 1, Lief. 1 and 2, 1922; C. R. Metcalfe and L. Chalk, *Anatomy of the Dicotyledons*, 2 vols., 1950.

## Plant classification

The phase of plant taxonomy concerned with the systematic arrangement of plants according to their relationships. Plants may be classified in many ways—by similarity of parts, complexity of structure, means of reproduction, or by combinations of these and other characteristics. However, the botanists of almost every country, except Great Britain, use the Engler-Prantl system of plant classification or modifications of this fundamental system. This system, set forth in *Die natürlichen Pflanzenfamilien*, by the German botanists, A. Engler and K. Prantl, is dominant in a majority of the large herbaria and published floras. Based primarily on natural relations, the Engler-Prantl system employs the following categories.

The basic unit of taxonomic work is the species, which is a grouping of individuals having essentially the same structure and life history. When a number of different species are found to have certain fundamental characteristics in common, they are grouped into a larger category called a genus. In the same manner, on the basis of inherent similar characteristics, related genera are grouped into families, families into orders, and orders into classes. The classes of the plant kingdom are frequently arranged into twelve different phyla. Each phylum represents one of the largest divisions of the plant kingdom. Its members have fewer characteristics in common than are found in the families or any of the lesser categories. For ex-

ample, a vascular system (specialized food- and water-conducting tissue) is the one main characteristic of all the members of the phylum, Tracheophyta. All the major classification groups may be broken down into smaller categories designated by such terms as subkingdom, subphylum, suborder, subfamily, tribe, and subgenus. Species may also be subdivided into such smaller categories as subspecies (ssp.), variety (var.), subvariety (subvar.), form (f.), and clone (cl.). The name of an order usually ends in *-ales* (Rosales); a suborder, in *-ineae* (Rosineae); a family, usually in *-aceae* (Rosaceae); a subfamily, in *-oideae* (Caesalpinoideae); and a tribe, in *-eae* (Pomaceae). The term taxon is used to designate any category whatever its rank: species, genus, family, or order. However, no one of these categories can be defined precisely. Each botanist fixes the limits according to his own views. See PLANT KINGDOM. [P.D.S.]

**Bibliography:** See PLANT TAXONOMY.

## Plant community

A plant community is an association of plants. Plants of various species are found growing together as vegetation, and certain combinations of species are found repeated in homogeneous areas of similar ecology, or biotopes, so often that generalizations can be made concerning these combinations. A plant community, then, has a certain species composition. A list of the plants occurring in a stand can be made by species names and by life forms. A list of all species is desirable; usually only vascular plants, bryophytes, and lichens can be recognized in the field. It is often necessary to take herbarium specimens, and such vouchers will document the study permanently. Ordinary taxonomic nomenclature is usually used, but a constant effort to improve this and to split the species into biotypes of more uniform relationships to environments must be made. The species list is limited, because within a given community the rate of increase of species number with increasing area is inversely proportional to the area investigated.

**Characteristic species.** Some kinds of plants are characteristic of a particular species combination: they are found only in one kind of combination wherever they occur, or regionally, or perhaps locally. Other plants are always found in a particular plant community. Still other plants occur in several kinds of communities; some are almost ubiquitous. Advantage is taken of such facts to classify the plants found into characteristic species which are exclusive to a given kind of vegetation or always found in it, differential species which occur in only one of two related communities, and accompanying species which show little or no preferences. The value of a given plant community as an indicator of habitat is determined largely by its characteristic and differential species, and it is by these species that plant communities are recognized. See PLANT SOCIETIES.

**Properties of plant communities.** It is possible to arrange the lists of species and associated eco-

logical habitat data made for various stands of vegetation in other ways than into types of plant communities. They can be arranged to describe gradients, series, continua, or functions in correspondence with various habitat factors. Properties of the vegetation other than species composition can be used to help characterize plant communities. Total yields per unit area, such as tons of forage/hectare or cubic meters of wood/hectare, life forms, dispersal or pollination spectra, and total contents of certain chemical elements all can be used as properties of plant communities.

Plants and animals which are associated also form a community, a biocoenose. Usually the plant community forms the fixed substratum for the animals, which may be mobile. See BIOLOGICAL PRODUCTIVITY.

**Structure of plant communities.** Plant communities have a structure varying in complexity from a many-layered forest to a unistratal polar or a hot desert cryptogam community. Moss, low and tall herb, low and tall shrub, and several tree layers may be present, and there may be epiphytic societies on the tree trunks and branches. In complex communities the aspect changes throughout the year as various groups of plants go through the stages of their life cycles at different times. Given the species list for a stand of vegetation, it is possible to assume much about the structure of the plant community from knowledge of the species concerned. However, one reason for studying plant communities in addition to individual plants is that in various communities individual species behave differently. Thus fireweed, *Epilobium angustifolium*, occurs in many forest communities of the Northern Hemisphere. It is usually sterile, but it flowers and proliferates abundantly when the forest is destroyed by fire or cutting, producing a new habitat and opportunity for a new plant community. See SUCCESSION, ECOLOGICAL.

**Dynamics.** The functioning of plant communities is analogous to the physiological processes taking place in the individual plants of which the community is composed, but significant interactions between plants modify, for example, the water regime of forest floor plants and the carbon dioxide made available for photosynthesis by the green plants. The physiological tolerances of individual plants to features of their habitats are thus modified to ecological tolerances by competition with other plants in the community.

The relations of plant communities to environment are systematized under various factors of the environment. Thus, plants react to such features of the environment as regional climates, soil parent materials, topographical features as these condition local climates, ground water, wind, snow deposition, fire, and the biota available to the biotope concerned. Man is a most important part of this biota, both in uncivilized and in civilized states. Plant communities in different geographical regions differ perhaps first of all because the floras available in the different regions differ, even

though ecologically the regions may be quite similar. The combined and interacting effects of all these groups of factors produce at a given time a particular ecosystem or combination of plant community in its environment in which the vegetation and environmental properties stand in functional relationship to each other. Thus on a continental scale the change in regionally representative plant communities from the short-grass high plains of Colorado east through the tall grass of Kansas to the deciduous forest of Ohio can be interpreted as a reaction to decreasing moisture along a given annual isotherm, such as 11°C. In the mountain ranges of the western United States, transitions from shadscale (*Atriplex confertifolia*) desert to sagebrush (*Artemisia tridentata*) semidesert to oakbrush (*Quercus gambellii*) chaparral to spruce-fir (*Picea engelmannii*-*Abies lasiocarpa*) coniferous forest to alpine herbaceous vegetation are related to altitudinal changes in these continental climates. Precipitation increases from 100 to 1000 mm whereas temperatures drop from 10 to 0°C as annual means. See ECOSYSTEM.

**Other factors.** These regional changes in vegetation can be found when only climate changes; the other factors of the environment are fixed at some particular values. If they are fixed at another set of values, the sequence will be quite different. A modification of the relief factor in the Middle Western case, a shift from the well-drained uplands to river floodplains, will result in riverine forest communities of various types all along the isotherm. If temperature is drastically lowered, as in the Arctic even 100 mm of precipitation will result in bog vegetation. The sequence of plant communities which corresponds to a change in one factor of the ecosystem is a function of those other factors of the ecosystem which have been constant.

**Climax.** Static situations are described above. However, vegetation is dynamic; it evolves. Given a fixed set of the environmental factors operating on a bare area, this area will change in the types of plant communities it supports, at a constantly decreasing rate until a steady state is attained. These equilibrium stages are climax plant communities. At such an equilibrium, which seems to be reached in a few hundred years depending on the ecosystem, the effects of climate in determining the kind of vegetation often become paramount. Although the effects of climate may become paramount in many ecosystems, in others with extremes of one of the other factors, the effect of this latter factor may persist indefinitely. Thus, very coarse-grained, sandy soil parent material may continue to support a plant community quite different from that on the surrounding hard land, as in the Sand Hills of Nebraska with their tall grass vegetation surrounded by climax mid- and short-grass.

In addition to the short-term genesis mentioned above there are changes of the environmental factors themselves which result in historical changes in plant communities. Invasion of plants new to the flora, as the chestnut blight into the hardwood for-



est of eastern North America which almost totally killed one of the former leading dominants in this forest, or the postglacial climatic changes which have been so well documented by pollen analyses of bog sections, are examples. If one of the factors determining a plant community changes, it is an axiom of plant ecology that the community will change. See CLIMAX COMMUNITY.

**Distribution.** The distribution of plant communities over the face of the earth has been studied more from a physiognomic than from a floristic viewpoint. Repetitions of physiognomically similar types of vegetation do occur in widely separated parts of the earth with similar climates. The ever-green sclerophyll, chaparral, of winter-wet, summer-dry, mild climates in the Mediterranean region of Europe, Australia, South Africa, California, and Chile is an example of floristically completely diverse regions having at least superficial similarities in the appearance of plant communities because of their similar structure. See CLIMAX PLANT FORMATIONS.

**Classification.** Finally, plant communities may be classified. The most widespread system is that developed by J. Braun-Blanquet and used extensively in Europe. Floristically similar stands of vegetation with some characteristic species in common are abstracted into associations denoted by the terminus -etum. Associations are combined into alliances (-ion), these into orders (-etalia), and these into classes (-etea). Classes in general coincide with broad, physiognomically defined kinds of vegetation or formations. The next higher unit is a floristic one recognizing such differences as those between the Mediterranean flora and that of central and northern Europe. Obviously, if two regions have different floras, they must also have different plant communities. See ANIMAL COMMUNITY; COMMUNITY. [J.M.A.]

**Bibliography:** J. Braun-Blanquet, *Pflanzen-soziologie*, 1951; H. Ellenberg, *Grundlagen der Vegetationsgliederung*, 1956.

## Plant disease

A great obstacle to the successful production of cultivated plants, plant disease is also sometimes destructive in natural forests and grasslands. Despite large expenditures for control measures, diseases annually destroy close to 10% of the crop plants in the United States, before and after harvest, resulting in a financial loss of at least \$3,000,000,000.

Diseases may destroy plant parts outright by rotting, or may cause stunting or other malformations. Most diseases are caused by parasitic microscopic organisms such as bacteria, fungi, algae, and nematodes or roundworms, although a few are caused by parasitic higher plants, such as dodder and mistletoe. Many are caused by viruses, and some are caused by poor soil conditions, unfavorable weather, or by harmful gases in the air.

The living organisms and viruses which cause disease are called pathogens. Most pathogens can

multiply extremely rapidly, the bacteria by simple division, the fungi by producing spores which behave as seeds but are much smaller and simpler in structure. Bacteria are about .0005 in. long, and fairly large fungus spores about .0001 in. Virus particles are not visible with ordinary microscopes; they can multiply a millionfold in a short time. Roundworms reproduce by means of eggs.

Most pathogens can be disseminated quickly and widely by wind, water, insects, man, and other animals. They infect plants through wounds, pores (stomata), or by penetrating plant surfaces. Each kind of pathogen can attack only certain kinds of plants or plant parts. Once inside the plant, living pathogens obtain their nourishment from it in various ways, destroying plant tissues or weakening the plant by robbing it of its food substances. The rapidity of growth and reproduction of pathogens and of disease development varies with the kind of pathogen and host and with soil and weather conditions. Some pathogens thrive best in hot weather, others in cool weather. Extensive and destructive epidemics develop when all conditions favor the most rapid development of the pathogen.

Good cultural practices, chemical disinfestation of planting materials, spraying or dusting with appropriate chemicals to protect against air-borne infection, and the use of resistant varieties are the principal control measures.

Discussed in the following sections are the economic importance, nature, and causes of plant diseases; the characteristics, growth, and reproduction of pathogens; the infection stage and development of diseases; the dissemination of pathogens; and the diseases to which plants are subject in storage. Discussion of other aspects can be found under separate titles or under the names of plants infected.

**Economic importance.** All plants and their parts are subject to diseases which may be caused at various stages of their life cycles not only by microorganisms, but also by higher plants, injurious salts in the soil, and harmful gases in the air. Diseases may rot the seed, kill plants, or make them poor and unsightly; they may cause root rots, stem cankers and rots, leaf spots and blights, blossom blights, and fruit scabs, molds, spots, and rots. In transit and storage they cause rots of fleshy fruits and vegetables; mold sickness of wheat, rice, corn, and other grains; and discoloration or rotting of wood and wood products.

When weather favors their development, some diseases become epidemic and ruin vast acreages of economically important plants. The historic potato famine in Ireland in the 1840s, resulting in the death of 1,000,000 people, was due to epidemics of potato late blight. Chestnut blight has ruined the chestnut forests of the United States. Stem rust destroyed about 300,000,000 bushels of wheat in the United States and Canada in 1916. In the United States it destroyed 60% of the spring wheat in 1935, and 75% of the macaroni wheat and 100% of the spring bread wheat in both 1953 and 1954.

stem rust, only one of more than 3000 kinds of plant rusts, has been similarly destructive in other wheat growing areas of the world, and it continually menaces wheat, oats, barley, rye, and many grasses. The *Helminthosporium* disease of rice was the principal cause of a famine in which a million or more people died in India in 1943.

Plant diseases reduce potential U.S. crop production by about 7% a year, equivalent to production on about 25,000,000 acres. This loss does not include costs of control measures, such as sprays and after-harvest losses. Spraying potatoes to control late blight often costs \$35 an acre; spraying fruits is similarly costly, and disease and pest control constitutes about 50% of the field costs in producing bananas.

Plant diseases are a dangerous threat to man's future subsistence. Much of the world is now underlarded and acute food shortages often occur in many areas. The situation tends to become worse as population increases by many millions each year. Plant diseases, old and new, are a critical limiting factor in food production. The degree to which they can be controlled will help determine whether the world can feed its rapidly growing population. [FCSN]

**Nature of diseases.** In the broad sense, disease in plants may be considered as any physiological abnormality which produces pathological symptoms, reduces the economic or aesthetic value of plant products, or kills the plant or any of its parts. Damage caused by wind or lightning or predation of insects or other animals, is not usually called disease, although such injury to living plants may result in a physiological disturbance which is truly disease. Decay of storage organs like tubers and roots is disease because such plant parts are living. Decay of lumber is disease only by extension of the definition, although the processes may be similar.

Disease in plants is usually evidenced by abnormalities in appearance, called symptoms, or by the presence of a pathogen in or on the plant. Some diseases, however, have no obvious symptoms; potato virus X, for example, reduces the yield of potatoes without apparent changes in the appearance of the plants.

The symptoms of plant diseases may be death (necrosis) of all or any part of the plant, loss of turgor (wilt), overgrowths (hypertrophy and hyperplasia), stunting (hypoplasia), or various other changes in the structure and composition of the plant. Necrosis may affect any part of the plant at any stage of growth. A rapid death of foliage is often called blight (Fig. 1), whereas localized necrosis results in leaf spots and fruit spots. Necrosis of stems or bark results in cankers (Fig. 2). Wilting may be slow or rapid, and it is usually more pronounced in dry than in moist soil. Necrosis eventually follows persistent wilting. Overgrowths composed primarily of undifferentiated cells are called galls (Fig. 3), the term tumor being less commonly used to designate these struc-



Fig. 1. Common bacterial blight of bean (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)



Fig. 2. Southern bacterial wilt of tomato. The plant shows leaf epinasty and wilt (After Kelman from J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)



Fig. 3. Crown gall of apple, (A. J. Riker from J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

tures. A bunch of small, abnormal shoots is often referred to as a witches'-broom. Underdevelopment or stunting may affect the entire plant or only certain of its parts.

Chlorosis (lack of chlorophyll in varying degree) is the most common nonstructural evidence of disease. For example, in leaves it may occur in stripes or in irregular spots (mosaic). Various degrees of curling and crinkling of the foliage generally accompany chlorosis. Sometimes there is also other abnormal coloration such as shades of red and brown.

A number of diseases may cause similar symptoms. These may be characteristic enough to permit diagnosis, but often it is necessary to identify the causal organism for exact diagnosis.

**Causes of plant diseases.** Usually two or more causes operate simultaneously to produce plant disease. For example, if a parasite is involved, the weather will influence the growth of the parasite as well as the plant's susceptibility to the parasite. The following subsections describe the influence on plant diseases of animals, plants, and viruses; soil conditions; weather; agricultural practices; industrial by-products; and plant metabolism.

**Animals, plants, and viruses.** Nematodes and insects are the animals that most commonly cause plant disease (Fig. 4). Although herbivorous animals, including many insects, bite off and swallow plant parts, the parts removed are not diseased and the animals are predators, not pathogens. However, the loss of the parts eaten may cause the rest of the plant to become diseased. Conversely, some insects are true pathogens because they remain on or in the plant and cause disease symptoms typically associated with the insects involved. Such



Fig. 4. Nematode galls incited by *Meloidogyne* sp. (a) On tomato. (b, c) On parsnip. (After Cox and Jeffers from J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

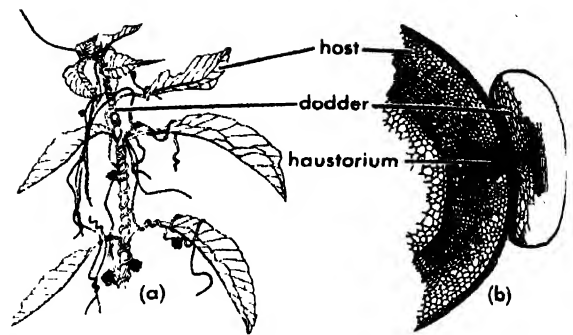


Fig. 5. (a) Dodder attached to host. (b) Section through host and showing haustorium of dodder extending into the host. (From F. W. Emerson, *Basic Botany*, 2d ed., Blakiston, 1954)

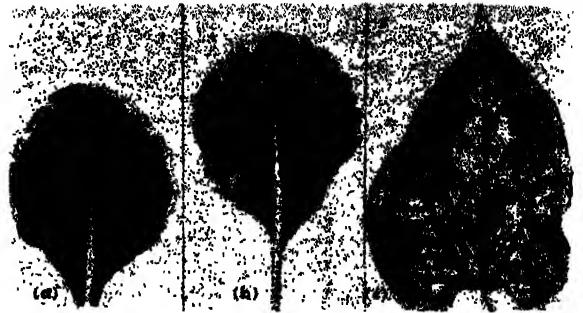


Fig. 6. (a) Potassium-deficiency disease of cabbage. (b) Iron-deficiency disease of cabbage. (c) Magnesium-deficiency disease of bean. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

symptoms may include yellowing, leaf curl, and overgrowths. Many nematodes are true parasites, hence pathogens, since they cause rots, overgrowths, and other plant abnormalities (see INSECTA: NEMATODA).

Certain algae, fungi, and bacteria are plant pathogens that cause disease. Most plant diseases are due to fungi; less than 200 are known to be caused by bacteria, and even fewer are caused by algae and parasitic seed plants such as dodder and mistletoe (Fig. 5).

Many plant diseases are caused by viruses, which are neither plants nor animals but behave like living things in many ways and may properly be called pathogens (see PLANT VIRUS).

**Soil conditions.** Deficiencies of mineral nutrients in the soil are a frequent cause of plant disease (Fig. 6). Often the deficiency can be identified by characteristic plant symptoms. For example, yellowing of the leaf tip and midrib of corn indicates nitrogen deficiency; yellowing of the margins, potassium deficiency. However, the symptoms may vary somewhat in different plant species. In addition, deficiency diseases may be difficult to diagnose, since they sometimes resemble those caused by viruses.

Besides nitrogen, potash, and phosphorus, which plants need in relatively large amounts, smaller quantities of sulphur, calcium, and ma-



Fig. 7. Boron-deficiency disease of garden beet. (a) Internal necrosis of tissue in the secondary cambial rings. (b) Leaves become stunted, and dormant buds of the crown are stimulated but form small distorted leaves. Internal necrosis near the exterior of the root leads to collapse of the outer tissue to form cankers. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)



Fig. 8. Blackheart of potato. (From J. C. Walker, *Plant Pathology*, McGraw-Hill, 1950)

nesium are required. Boron, iron, copper, manganese, molybdenum, zinc, and other minerals are used in such minute amounts that they are called trace elements. However, if one of the latter is missing, a typical disease may result, such as dry rot of rutabagas, which is due to boron deficiency (Fig. 7). See PLANT, MINERALS ESSENTIAL TO.

Frequently deficiencies of minerals cannot be determined by soil analysis alone, because the minerals may be present in chemical combinations that plants cannot use. For example, iron is often unavailable on high-lime soils, even if it is present in the soil in appreciable quantities (see PLANT, MINERAL NUTRITION OF).

Besides lime, excess amounts of many other chemicals may be present in the soil and cause plant disease. Excess of soluble salts causes "alkali injury" and aggravates drought damage; too much nitrogen may stimulate abnormal growth; while an excess of boron may cause necrosis and stunting. Unfavorable chemical balance in the soil may also result in excess acidity (low pH) or alkalinity

(high pH), either of which may inhibit normal plant growth (see PLANT GROWTH).

The soil is the principal source of water, which all plants need in varying amounts, depending upon the species. Too little available water slows growth and, below certain limits, results in wilting. Plants can recover from limited (transient) wilting, but if it is prolonged the affected parts die (see PLANT, WATER RELATIONS OF). Evergreens are often damaged by moisture loss on warm, windy days in winter, when the soil is frozen and they cannot absorb sufficient water to replace the loss.

Conversely, too much water in the soil results in oxygen deficiency, which will cause suffocation of the tissues of roots and other underground parts of most plants. Water-inhabiting plants, such as rice, are exceptions. Excess water in the soil favors certain kinds of fungus and bacterial diseases, and these are often confused with the purely physiologic effects of too much water.

High soil temperature during the growing season may also cause disease, for instance, internal necrosis of potato (Fig. 8).

The structure of soil (particle size and organic content) determines its water-holding capacity and hence affects both the conditions mentioned above and the ease with which plant roots penetrate the soil (see SOIL).

*Weather conditions.* Wind, lightning (Fig. 9), and hail may injure plants and cause true diseases such as those resulting from unfavorable temperatures (Fig. 10). Temperature effects range from



Fig. 9. Lightning injury of cabbage, as seen several weeks after the injury occurred. (a) Callus tissue on the stem at the ground level where the charge entered the plant. (b) Interior of plant shows in (a). The paths whereby the charge passed through the cortex and the vascular ring are evident. The pith was killed, and as the tissue collapsed, adventitious roots formed in the cavity. (c) Dormant buds stimulated to growth at leaf axes just below where the charge entered. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

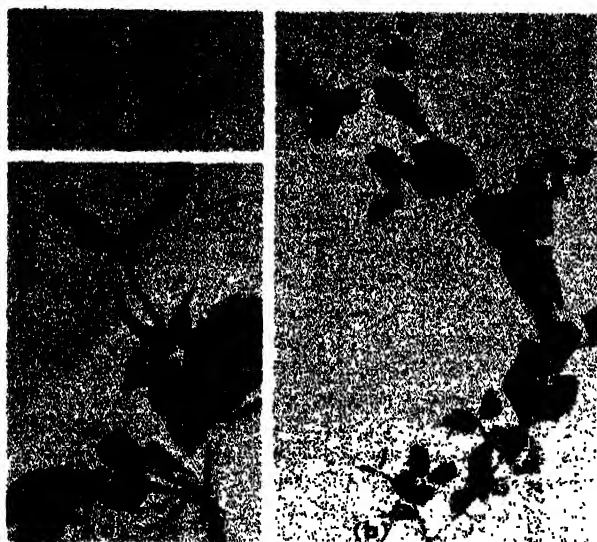


Fig. 10. Freezing injury of pea. Symptoms which are apparent several weeks after injury. (a) Enlargement of the injured growing point in (b); in the youngest leaf the stipules and the first pair of leaflets have assumed abnormal shapes and the second pair of leaflets did not form. (b) Following killing of the growing point at the left, a lower dormant bud grew out to form the main stem. (c) Necrotic bands in a pair of leaflets which were developing at the time of injury. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

poor development of plants grown in climates too cold or too warm to actual frost or heat damage. For example, tomatoes grow poorly and drop their blossoms in cool weather; direct sunlight on the fruit may kill the tissue, causing sunscald; and the foliage is severely damaged by even light frosts that would not harm cabbage.

Exposure to gradually decreasing temperatures in the autumn hardens perennial plants such as fruit trees against winter cold, but exposure to the sun in the late winter may make the bark tender again. This tissue is killed when it freezes at night, causing another kind of sunscald.

Although high temperatures may literally cook plant tissue, with such results as "heat canker" of young flax and sunscald of tomato fruit, the commonest effect is to increase water loss by transpiration, resulting in drought damage. Wind has the same effect, the degree depending upon its velocity and relative humidity.

In most green plants deficiency of light causes weak, spindly growth, and chlorosis, although some species can endure much shade. House plants are frequently affected in this manner, but excess shading by buildings or other plants will produce the same effect out of doors.

**Agricultural practices.** Mismanagement of soil, including untimely applications of irrigation water and fertilizer, can cause plant disease, but other agricultural practices are frequently injurious. The more common of these injuries result from the improper use of chemicals such as fungicides, insecti-

cides, and herbicides (see AGRICULTURE; FUNGICIDE; HERBICIDE; INSECTICIDE).

Nearly all fungicides are injurious to plants as well as to fungi, although the damage to the plants is usually much less than the potential injury from the diseases controlled by the fungicides. Examples of effects are increased transpiration caused by Bordeaux mixture on tomatoes, russetting of fruits caused by lime sulphur on apples, and yield reductions without visible symptoms caused by other chemicals. Conversely, the fungicide may contain a nutrient, such as zinc, that is deficient in the soil, and much better growth of the plant may result.

Chemicals used for seed treatment are frequently toxic, especially to some species of plants. For example, plants of the cabbage family are stunted by copper-containing seed treatment materials. Vegetative organs, like potato tubers, are very susceptible to chemical injury, and strong poisons like mercuric chloride often do more harm than good. Materials applied to the soil to control fungi, bacteria, and nematodes may injure plants grown in the soil too soon after treatment.

Some crop plants are very sensitive to herbicides, being affected by very minute amounts of such things as 2,4-D (2,4-dichlorophenoxyacetic acid). Tomatoes may be affected from sources far removed. Symptoms of 2,4-D are sometimes confused with those of virus diseases.

**Industrial by-products.** The fumes from ore smelters frequently cause widespread symptoms of plant disease, including stunting, yellowing, and necrosis. Where atmospheric inversion layers prevent their escape, even traffic and domestic fumes may be toxic (see ATMOSPHERIC POLLUTION).

**Plant metabolism products.** Brown areas on stored apples (scald) may be caused by ethylene gas produced by the apples (Fig. 11). This gas occurs in small quantities in many healthy plant tissues but is produced in greater amounts by diseased and aging cells. Ethylene gas may also cause yellowing in plants, and it accelerates ripening in certain fruits such as banana (see PLANT METABOLISM).

#### PLANT DISEASE PATHOGENS

Most pathogens are grouped primarily on the basis of their structure; but bacteria, being morphologically simple, are classified to a considerable extent by physiological characters. Viruses repre-



Fig. 11. Apple scald. (USDA)



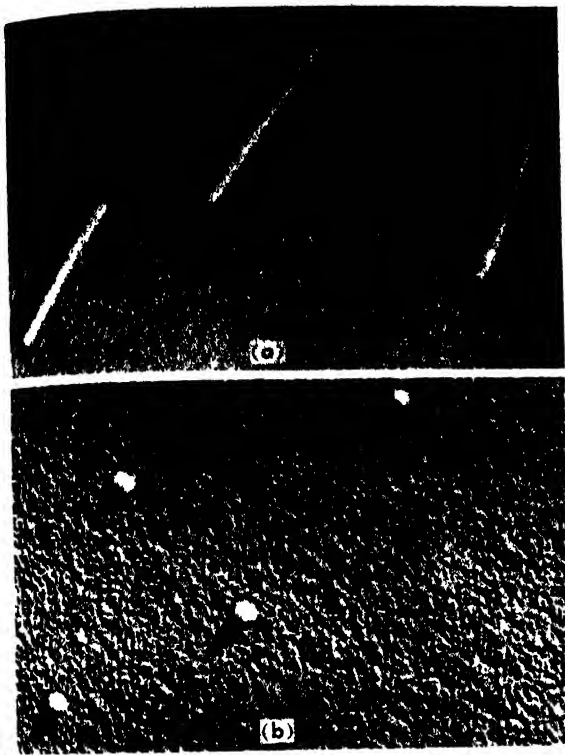


Fig 12. (a) Rod-shaped particles of the tobacco mosaic virus (b) Polyhedral particles of the squash mosaic virus. Electron micrographs of preparations made by the freeze-drying technique. Magnification of approximately 100,000. (Paul Kaesberg from J C Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

sent a special problem, and such considerations as means of transmission and host symptoms are used in classifying and naming them.

Fungi, bacteria, and a few seed plants are heterotrophs; that is, they lack chlorophyll and consequently are dependent, directly or indirectly, upon green plants (autotrophs) for carbohydrates (see PHOTOSYNTHESIS). Animals and some fungi and bacteria are also dependent upon other organisms for nutrients such as amino acids and vitamins. Viruses seem to become an intimate part of the chemical make-up of the host plant (Fig. 12).

Plant pathogens usually penetrate into the host plant and grow within or between the cells (Fig. 13). Viruses are usually intracellular, and some are confined to the phloem, whereas plant pathogenic bacteria are usually intercellular or occur in the xylem (see PLANT ANATOMY). Fungi are composed of microscopic tubes called hyphae, by means of which plant pathogenic species penetrate into or between the host cells (Fig. 14). The powdery mildew fungi grow principally outside of the plant but send special absorptive organs (haustoria) into the host cells (Fig. 15). Some intercellular species of fungi also produce haustoria. Pathogenic seed plants, such as mistletoe and dodder, usually penetrate the host by means of rootlike absorptive organs. Pathogenic insects and nematodes may be wholly within the plant, or they may

remain superficial and penetrate the host with specialized mouth parts.

Most plant pathogens are parasites. Some, such as the rusts and powdery mildews, are obligate parasites, that is, can grow only on a living host plant. Viruses are also in this category, although they are not typical organisms. Fungi and bacteria that can use only nonliving food sources are called saprophytes.

Most of the fungi and all plant pathogenic bacteria can grow on nonliving organic matter as well as parasitically on living matter; these are called facultative saprophytes. Some organisms live primarily as saprophytes but also have the ability to parasitize weakened plants and are therefore called facultative parasites. Many plant pathogens have both a parasitic (or pathogenic) and a saprophytic phase of development.

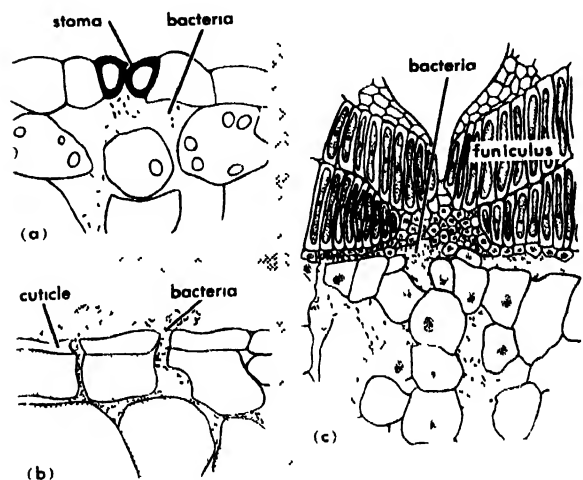


Fig. 13. Common bacterial blight of bean. (a) Invasion through stomata. (b) Invasion through rift in the cuticle of the cotyledon. (c) Invasion of the seed through the tissue of the funiculus. (After Zaumeyer from J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

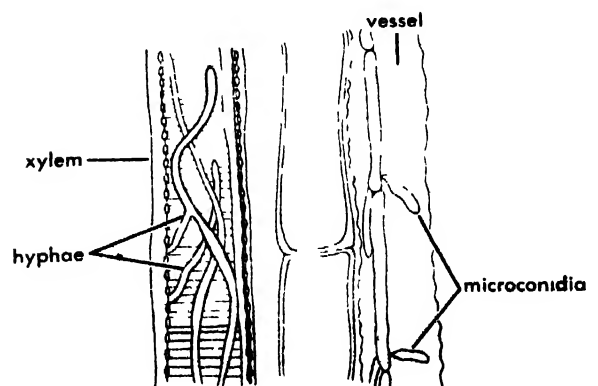


Fig. 14. The cabbage-yellows organism in tracheae of the cabbage plant. Note formation of microconidia in the vessel. (After Gilman from J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)



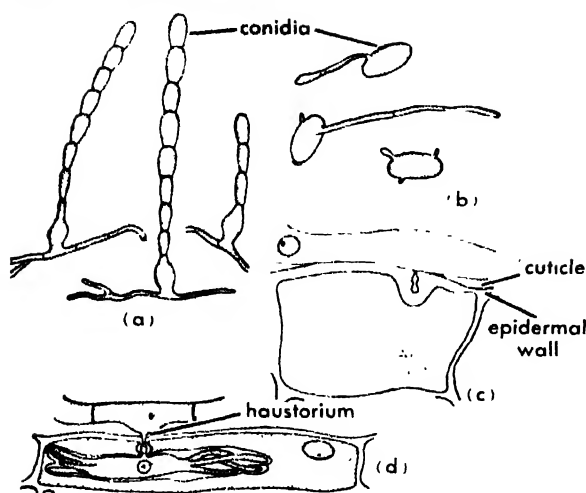


Fig. 15. *Erysiphe graminis*. (a) Conidiophores and conidia (spores). (b) Germinating conidia. (c) Penetration of cuticle and epidermal wall. (d) Haustorium. (a, b, after Reed; c, d, after G. Smith from J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

**Symbiotic relations of organisms.** Parasitism is the one of a series of associations characterized by intimate physical union of taxonomically dissimilar organisms. Such relationships are known as symbiosis, and may be neutral, beneficial, or harmful to the symbionts. An association such as that of legumes and nodule bacteria, beneficial to both partners, is called mutualistic symbiosis. Parasitism is antagonistic symbiosis.

There are different degrees of parasitism. In the early stages, the association between rust fungi and their hosts may appear to be almost neutral, harming the plants little. Other fungi, such as those rotting fruit, can become established only in dead tissue, producing enzymes or toxins that kill adjacent living cells which they then inhabit. Some biologists say that such organisms are saprophytes, not parasites, because they never colonize living host tissue. But the term parasitism is generally used to refer to a relationship with the host plant as a whole, because the degree of intimate relationship is often difficult to determine.

**Ecologic relations of organisms.** Associations of organisms in the same environment without physical union are called ecologic and are often very important in plant disease (see *ECOLOGY*). As in symbiosis, the effects may be beneficial, neutral, or harmful. Metabiosis occurs when one organism uses a substance for food and produces a by-product that enables another to grow. If the benefits are reciprocal, the relationship is called synergism, as when the fungus *Mucor ramannianus* produces pyrimidine and *Rhodotorula rubra*, a non-sporulating yeast, makes thiazole (see *CRYPTOCOCCALES*; *PHYCOMYCETES*). These chemicals are components of thiamine, which both organisms need but which neither can produce alone. If deleterious substances (antibiotics) are produced, the relationship is called antibiosis.

All of these relationships may be important to the survival of certain plant pathogens, especially some of those which live in the soil part of the time. Metabiotic and synergistic relationships may help them to survive; antagonistic relationships will hinder survival. One of the goals of the plant pathologist is to encourage antibiosis that will eliminate certain soil-inhabiting pathogens.

Ecologic associations may exist between two or more pathogens inhabiting the same host plant as a common environment. When fire blight bacteria parasitize apple twigs and permit the entrance of canker and wood-rotting fungi, the relationship between the bacteria and the fungi is metabiotic. The molds *Oospora citri aurantii* and *Penicillium digitatum* can rot fruit more rapidly together than either can alone. This is synergism. Antagonism seems to exist between races of the potato late blight fungus, and one will replace the other when they parasitize a potato plant together.

Even the relationship of host and pathogen may be ecologic at first. For example, *Rhizoctonia solani* in the soil causes visible injury to the roots of soybean before touching them. Accordingly, the fungus is at first antibiotic to soybean; later it becomes parasitic and pathogenic. [C.J.F.]

**Growth and reproduction.** Many plant pathogens, especially among the bacteria, fungi, and viruses, can multiply with amazing rapidity under favorable conditions. Viruses, although not generally considered living organisms, may increase a millionfold a few days after introduction into the right place in the right kind of living plant, when temperature and other environmental conditions are favorable to the virus.

**Food requirements of bacteria and fungi.** Although lack of chlorophyll prevents these organisms from using solar energy to synthesize basic carbohydrates from carbon dioxide and water as green plants do, their basic nutrient requirements are essentially the same as those of higher plants. They require carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur as structural elements. In addition, they need the metallic elements potassium, magnesium, iron, zinc, copper, calcium, gallium, manganese, molybdenum, vanadium, and scandium. Potassium and magnesium, needed in relatively large amounts, are designated macroelements; the others, some of which are needed in minute amounts, are often designated microelements. Vitamins, enzymes, and hormones are also needed for growth and reproduction.

For experimental purposes, pure cultures of facultative saprophytes are grown in the laboratory on sterilized synthetic media containing sugars or some other source of carbon, salts of the other necessary elements, and essential vitamins for those organisms which cannot synthesize their own. Natural plant products, such as potato broth, steamed cornmeal, or oatmeal, often are used as nutrient bases. Liquid media are used for some purposes; for others, the nutrient solutions are solidified with gelatin or agar. Nutrient requirements for growth

and reproduction are best determined by varying the composition of synthetic media. Studies on the effects of temperature, light, and other environmental factors are facilitated when organisms can be grown on culture media. Although all pathogenic organisms have some requirements in common, they differ greatly in special requirements, both on artificial media and on host plants. By growing pathogens artificially, much is learned about them which enables the development of better control measures.

**Host selectivity of bacteria and fungi.** Among the approximately 150 species of pathogenic bacteria and the many thousands of fungi, there are wide differences with respect to the kinds of plants and plant parts on which they can grow. Flax rust (*Melampsora lini*) grows only on wild and cultivated flax, asparagus rust (*Puccinia asparagi*) principally on asparagus; *Xanthomonas campestris*, the bacterium which causes black rot of cabbage, cauliflower, and related plants, parasitizes members of the mustard family only. On the other hand, the fungus *Rhizoctonia solani* causes root rot of potatoes, alfalfa, clover, and hundreds of other species in many different plant families; the bacterium *Agrobacterium tumefaciens* causes crown gall on grape, raspberry, chrysanthemum, and numerous other plants; the bacterium *Erwinia carotovora* causes soft rot of almost all kinds of fleshy vegetables.

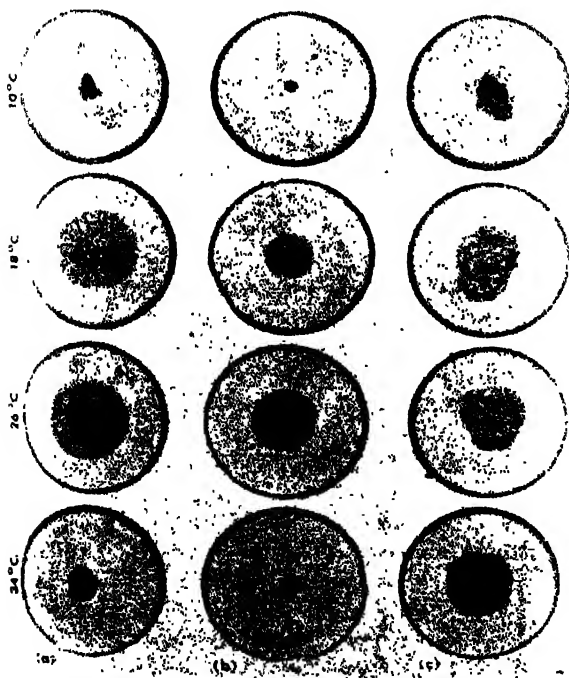


Fig. 16. Effect of temperature on rate of growth of different mutant lines of corn smut (*Ustilago maydis*). Note difference in ability of lines to grow at the extremes: (a) line W.Va. A8-3-1 is intermediate; (b) line W.Va. A8-5-4 grows scarcely at all at the extremes; (c) line W.Va. A8-5-2-1 grows fairly well at 10°C. (Univ. Minn. Agr. Exp. Sta. Tech. Bull. 65)

Some pathogens attack only a few plant parts or tissues, others attack many. Some attack roots only, others attack stems, others cause leaf spots, and still others attack fruits. Some attack young tissues, others attack old ones. There are diseases of youth and of age, of herbaceous plants, and of woody plants. Some pathogens parasitize all plant parts of susceptible hosts at all stages of development. To understand and control the numerous diseases of thousands of kinds of plants, it is necessary to learn the conditions under which each pathogen thrives.

**Environmental factors.** The rate and kind of growth and reproduction of pathogens are affected by nutrition, moisture, temperature (Fig. 16), light, the acidity or alkalinity of the medium, the relative amounts of oxygen and carbon dioxide, and by other microorganisms with which they must compete. Most pathogens require free moisture for germination and infection, although some powdery mildews can germinate in dry air. Soil moisture sometimes is a determining factor in growth and reproduction. Some pathogens that live in the soil thrive best at high moisture content, some at low. Temperature determines the geographical and seasonal occurrence of many diseases, since the cardinal temperatures—the minimum, optimum, and maximum—differ for different pathogens. The peach leaf curl fungus, the potato late blight fungus, and yellow rust of wheat develop best at a relatively low temperature; the peach brown rot fungus, the potato wilt and brown rot bacterium, and stem rust of wheat develop best at a relatively high temperature. Light has less influence than temperature on the growth of pathogens in nature, but it strongly affects reproduction of some fungi. Some soil organisms, such as the potato scab bacterium, like an alkaline (high pH) soil; some, such as the cabbage clubroot fungus, like an acid soil (low pH).

**Reproduction of bacteria and fungi.** As far as is now known, plant pathogenic bacteria reproduce only by simple fission. A single bacterium divides into 2, the 2 into 4, and so on. As division may occur every 20 to 30 minutes, a single bacterium could produce a progeny of 300,000,000,000 within 24 hours. The rate, however, varies with the kind of bacterium, with its nutrition, with temperature, and with other environmental conditions.

Most fungi, however, reproduce both asexually and sexually. In many of them asexual reproduction results in rapid multiplication (Fig. 17), whereas sexual reproduction results in the production of spores that can survive unfavorable conditions. In general, fungi continue to grow and produce asexual spores while the environment is favorable and nutrients are easily available; but they tend to produce sexual spores when growth is checked. Thus an asexually produced urediospore (summer spore) of wheat stem rust (*Puccinia graminis* var. *tritici*) can cause infection, the resulting mycelium grows for a time, and then forms a pustule containing 50 to 400,000 new urediospores.



Fig. 17. Spore-producing branches of *Penicillium* similar to the one from which the drug penicillin is obtained. Chains of spores are produced on the ends of branches. (Univ. Minn. Agr. Exp. Sta.)

The time required is only about a week at 75°F, but it increases to a month at 50° and even longer as temperature decreases. Each new spore can cause a new infection, and this process continues, at a rate that varies greatly with temperature, moisture, and light, until the wheat starts to ripen or growth is otherwise checked. Then the winter spores (teliospores) are produced; these differ from urediospores in appearance and cannot normally germinate until they have been exposed to winter weather. The apple scab fungus (*Venturia inaequalis*) may produce many successive crops of asexual spores (conidia) on the fruit and leaves during the growing season. But it does not produce sexual spores until the following spring, on infected leaves that have fallen to the ground the previous autumn. Some fungi, such as the ergot fungus (*Claviceps purpurea*) produce sclerotia, bodies made up of densely interwoven hyphae, which may survive winters or other unfavorable conditions and then produce fruiting structures under appropriate conditions in the spring.



Fig. 18. Fruit bodies of a tree-inhabiting mushroom, *Schizophyllum*. Basidiospores are produced on the sides of the gills. In *Schizophyllum* the gills are split lengthwise; in dry weather they curl up, and so conceal and protect the surface on which spores are borne. (Univ. Minn. Agr. Exp. Sta.)

Special stimuli are sometimes necessary to initiate the formation of fruit bodies (Fig. 18); some fungi require the stimulus of light for fructification, although they grow well in darkness; some require special temperature; others require certain nutrients or vitamins.

How, where, when, and the rate at which fungi grow and reproduce depend on their inheritance and their environment. The inheritance determines the limits within which the behavior of each kind of fungus can vary, and the environment determines its behavior under particular combinations of conditions. [F. C. S.]

### INFECTION AND DEVELOPMENT OF DISEASE

Infection of plants by a pathogen terminates a series of events that begins with inoculation, which is the contact of a susceptible part of a plant with the inoculum. Inoculum is any infectious part of the pathogen, such as spores, bacterial cells, or virus particles. Typically, inoculation is followed by entrance into the host, and infection follows entrance. A plant is infected when the pathogen starts taking nourishment from it.

The time between inoculation and infection is the incubation period. Because it is often difficult to tell when infection occurs, the incubation period is usually counted as the time between inoculation and the appearance of the first symptoms of infection.

The probability that infection will follow inoculation depends upon the vigor of the inoculum, the duration of favorable environmental conditions, and the resistance of the host (see PLANT DISEASE CONTROL). Usually only a small part of the inoculum produced reaches a susceptible plant, and only a small fraction infects the plant. Consequently, most plant pathogens survive and are destructive partly because they produce fantastically large amounts of inoculum.

**The inoculum.** Inoculum of viruses and bacteria consists of the individual virus particles or bacterial cells, respectively; the inoculum of fungi may be spores, pieces of hyphae, or specialized structures, such as sclerotia. Pathogenic plants like dodder produce true seed, and nematodes produce eggs, both of which function as inoculum.

Bacteria and viruses produce billions of cells or virus particles in infected plants, and each new unit theoretically can infect another plant. Fungi produce spores on the surface of hyphal growth or in a variety of specialized structures which may be large, as the giant puffball, or almost invisible to the unaided eye (Fig. 19). Some of the spore-producing structures function over a considerable period of time and, like bacteria, produce prodigious amounts of inoculum.

Bacteria and viruses are somewhat restricted as pathogens by having no special means of liberating themselves from the host, although the bacteria may ooze out in sticky droplets. For dissemination or transmission these pathogens depend chiefly upon plant contact, insects, or man, although bac-

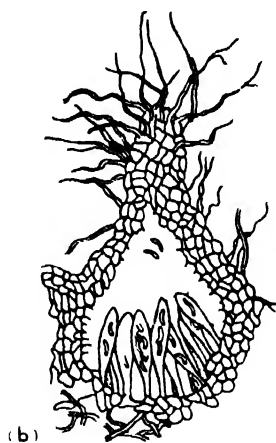
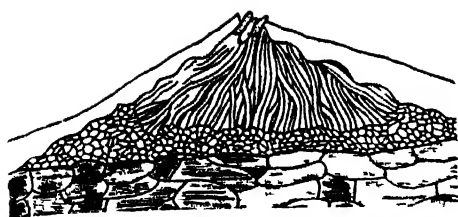


Fig 19 *Glomerella cingulata*. (a) Acervulus on apple fruit (b) Perithecium. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

teria may be spattered short distances by rain. Some fungi produce spores in sticky masses, like bacterial ooze, and are disseminated in much the same ways as bacteria. Other fungi have ways to liberate or forcibly eject spores into the air, where they can be carried by the wind. This gives fungus pathogens the potential of much farther and faster spread than the bacteria or viruses, although their arrival on a susceptible plant is much more a matter of chance than if insects carry the inoculum, because insects often seek similar plants for food (Dissemination is considered in greater detail in a later section of this article.)

Dormant inoculum is one of the most important, but not the only, means by which plant pathogens survive during periods when parasitic life is impossible. If the pathogen is within a perennial host, it is usually quiescent during the rest period of the host. Sclerotia and even the vegetative hyphae of some fungi may survive periods of drought and cold independently of the host. Other pathogens require the protection of the dead host plant, not so much against cold and drought as against antagonistic organisms (see ECOLOGIC INTERACTIONS). This is especially true of plant pathogenic bacteria, few of which survive long if separated from host tissue. Some viruses can live only minutes apart from the living host; others, like tobacco mosaic virus, remain infective for years in dried leaves.

At the beginning of the growing season, the first inoculum of a pathogen is called primary inoculum; that which is produced later on infected

plants, secondary inoculum. The primary inoculum of fungi may be resting spores or the surviving hyphae or sclerotia; often the hyphae or sclerotia produce spores which function as primary inoculum.

Many fungi produce two or more kinds of spores. Those formed late in the growing season (resting spores) usually will not germinate until after a period of dormancy and will survive more cold and drought than spores produced during the growing season. Some, like the spores of the cabbage clubroot fungus and the chlamydospores of the onion smut fungus, stay dormant for several years, thus assuring the species of survival if susceptible hosts are not grown on the land for several seasons. Such diseases are difficult to control by crop rotation.

"Repeating" spores typically are morphologically distinct from the resting spores, and are produced in great numbers on diseased plants during the growing season. They usually germinate rapidly whenever environmental conditions are favorable. Before germination, repeating spores can survive for periods ranging from several hours to several weeks, depending upon the species. This determines largely how far and under what conditions a pathogen will spread during the growing season.

**Spore germination.** Germination, as applied to spores or seeds, means the resumption of vegetative growth leading to the development of a new individual. In fungi this usually means the production of a hypha, called a germ tube (Fig. 20). Cell division of bacteria and the hatching of nematode and insect eggs are comparable processes so far as their function as pathogens is concerned.

Germination occurs if the spore is not dormant and if environmental conditions are favorable. This usually requires a certain temperature range and liquid water, although a few species of fungi (powdery mildews) germinate in humid air. Certain species also require the presence of food substances, special stimulants associated with the host, absence of inhibitors that may be produced by the pathogen, or certain degrees of acidity. Such requirements limit germination, but may be a benefit to the species. For example, the necessity

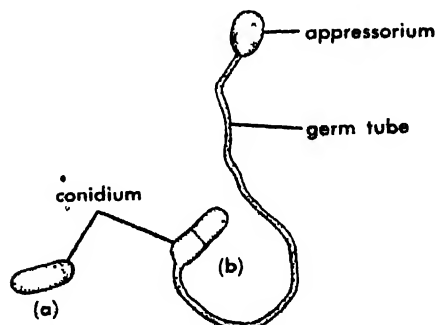


Fig. 20. *Glomerella cingulata*. (a) Conidium (spore). (b) Conidium that has become septate during germination; appressorium at tip of germ tube. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

for a host stimulant will prevent wastage of spores in the absence of the host.

When a nondormant spore is placed under favorable conditions, germination may follow in 45 minutes or only after several days, depending upon the species, age of the spores, and variations in the environment. Since conditions change rapidly, germination is a critical time for a fungus, because if it does not penetrate the host quickly the germ tube may be killed, especially by dryness. It is at this stage that fungi are most easily killed by fungicides.

**Establishment in the host.** For bacteria and viruses, entering a host is a passive process. Bacteria accidentally get into injuries or are put there by insects or other agencies; they may also be drawn by water into stomata, hydathodes, or necrotic areas. Viruses often are placed in the host by insects, but many can be transmitted when the sap from infected plants comes in contact with minute wounds in healthy plants.

Spores of fungi may also be carried into plants by various agencies, but many species have active means of penetration, the method usually being characteristic of the species. In some, germ tubes enter stomata by producing a flat structure (appressorium) over the stoma from which a hypha grows through the opening (Fig. 21). Others ignore the stomata; instead the appressorium adheres to the cuticle of the plant and forces a slender infection peg directly through the protective layer (Fig. 22). This apparently is accomplished entirely by pressure, as no enzyme action has been demonstrated.

Animal pathogens, like nematodes, have special mouth parts that pierce the plant, and the nematode may remain external or it may actually enter the plant.

Even after penetration, pathogens may fail to infect due to the presence of mechanical barriers, lack of proper nutrients, or the presence of inhibiting toxic substances. These factors depend not only on specific interactions between host and pathogen but also upon the environment. Successful establishment of the pathogen may mean killing the host cells and living upon the dead tissue, with or without the actual penetration of living cells. [C.J.E.]

**Development within the host.** After a pathogen has become established in a susceptible host, the rate of disease development under favorable conditions follows a sigmoid (s-shaped) curve with three major aspects: (1) the lag phase or incubation period, when infection is not evident externally; (2) the exponential phase, when the pathogen spreads rapidly in host tissues and symptoms and signs of disease appear; and (3) the senescence phase, when limiting mechanisms of either host or pathogen restrict further extension.

Disease development varies with genetic susceptibility of the host, genetic aggressiveness of the pathogen, and with many environmental factors that influence the host, the pathogen, and the inter-

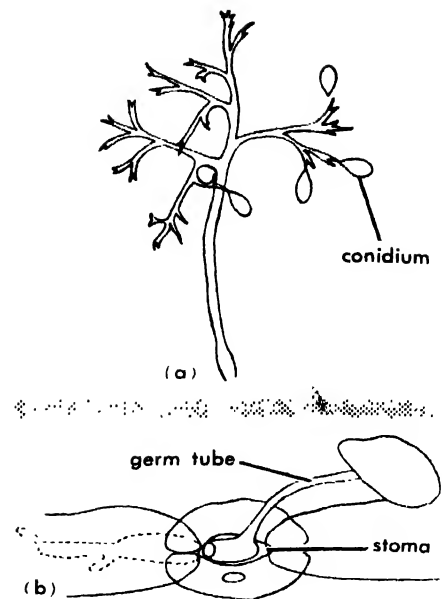


Fig. 21. *Peronospora destructor*. (a) Conidiophore bearing conidia. (b) Conidium germinating by a germ tube, the latter penetrating a stoma. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

actions between the two. Environmental factors influence the growth rates and the metabolism of the host and the pathogen; and the interrelations between these activities determine the pattern of disease development. Furthermore, the effects of past environmental conditions on the host may affect disease development, a condition known as predisposition when host susceptibility is increased. The combined effects of these factors on growth and development of healthy crop plants in nature are poorly understood, and the problem becomes increasingly complex when the plants become diseased.

Climate often determines the adaptability of plant species to geographic areas and may also determine the geographic distribution of their diseases. For each disease there are minimum, optimum, and maximum values for each critical environmental factor. The mean measurements of

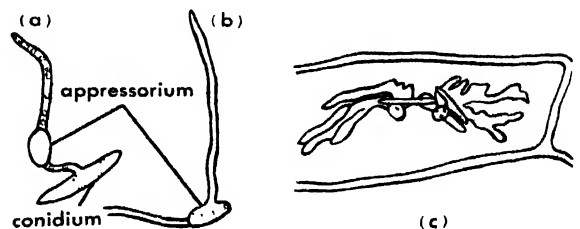


Fig. 22. *Colletotrichum circinans*. (a) Conidium which has germinated and formed an appressorium, which in turn has germinated. (b) Germinating appressorium. (c) Conidia and appressoria on the surface of the host and the subcuticular mycelium developing after penetration. (From J. C. Walker, *Plant Pathology*, 2d ed., McGraw-Hill, 1957)

weather, however, are often less important than the exact combinations of weather at critical times. Those environmental factors that deviate most from the optimum limit the development of a disease.

*Effect of temperature.* Temperature has a major effect on disease development and determines the seasonal and regional incidence of most diseases. For example, a succession of diseases attacks certain creeping bent grasses on golf greens in northern United States: snow mold occurs beneath melting snow in winter; red thread during the moderate temperatures of spring and fall; dollar spot during the warm temperatures of early and late summer; and brown patch during the hot weather of midsummer. The fungi causing these diseases may attack the leaves, grow into the crowns of the plants, and may kill the plants entirely in patches of characteristic size for each disease. These diseases produce similar effects, but temperature determines when each disease is most destructive.

The length of incubation period of disease is governed by prevailing temperatures. Many diseases, such as rusts, mildews, and leaf spots, cause only small lesions on above-ground plant parts. The damage to the plant depends on the number of lesions, which in turn depends on the number of disease cycles. The elapsed time from infection to spore production—the length of the incubation period—determines the frequency of disease cycles. Thus, temperature often determines whether pathogens can produce enough disease cycles for development of an epidemic. Temperature likewise may influence the symptom expression. Thus, symptoms of many virus diseases disappear or are masked at high temperatures. Temperature also determines whether certain wheat varieties are susceptible or resistant to certain parasitic races of stem rust. The effect of temperature on disease development may be principally on the pathogen, or it may be on the host. When the cardinal temperatures are the same for growth of the pathogen in culture and for development of the disease, the effect is principally on the pathogen. However, when the optimum temperature for growth of the pathogen in culture differs from that for maximum disease development, the temperature probably predisposes the host plant by weakening it.

*Effect of light.* Light affects disease development principally by its effect on photosynthesis and the assimilative processes of the host. Obligate parasites, such as rusts and powdery mildews, generally develop best when assimilation is maximal, although the severity of the disease lesion caused by some pathogens on some hosts may be decreased by high light intensities. Low light often weakens plants and thus predisposes them to diseases caused by facultative saprophytes.

*Effect of moisture.* Moisture is a major factor in germination and entrance of pathogens into the host. The moisture requirements of the established pathogen are supplied by the host, since the osmotic value (water absorption capacity) of the hyphae of the pathogen is always greater than that

of the parasitized host cells. Transpiration (water vapor loss) from diseased above-ground plant parts is greater than that of healthy parts. The water economy of the host is disrupted in wilt diseases by the effects of the pathogen on the translocation of water in the xylem and the osmotic permeability of foliage parenchyma, and in root diseases by the destruction of the tissues for water absorption and conduction. The rate of symptom development and death of the plant tissue in wilts and root rots is accelerated by excessively low atmospheric humidities and low soil moisture availability.

*Relation of soil.* Soil reaction, as regards hydrogen ion concentration, affects the development of many diseases in the soil. Potato scab is less severe in acid soils (below pH 5.2) while cabbage clubroot is not so severe in less acid soils (above pH 5.7). However, the extent to which the soil reaction affects the infectivity of these pathogens and the subsequent development of the diseases has not been determined. As the hydrogen ion concentration of the plant cell is relatively constant despite differences in the range of soil reaction, soil pH probably affects disease development indirectly by its effects on the availability to the host or pathogen of mineral nutritional elements in the soil.

Soil oxygen and carbon dioxide concentrations affect the development of root diseases. The effects on infectivity of the pathogen, on predisposition of the host, and on disease development have not been distinguished, although the development of the host is more adversely affected by high carbon dioxide and low oxygen tensions in the soil than is the growth of many fungal pathogens.

*Effects of nutrients.* The effects of nutrients are largely indirect since plants and their pathogens require the same essential mineral elements. However, the available amount of each mineral element and the balance between them affect the structure and physiology of the host and thus may be either favorable or unfavorable to the development of different pathogens. The principal mineral elements in fertilizers (nitrogen, phosphorus, potassium, and calcium) have the most pronounced effects. Diseases caused by obligate parasites such as rusts, powdery mildews, and many viruses develop best in "normal" plants having optimal mineral nutrition; while subnormal plant development due to inadequate or unbalanced mineral nutrition favors the development of many diseases, such as root rots, that are caused by facultative saprophytes. Some vascular pathogens are affected directly by the concentration of nitrogen compounds in the conductive tissues of the host.

**Plant disease epidemics.** Epidemics of plant diseases occur when a high percentage of the host population in a certain area is affected with sufficient severity to limit either growth, survival, or crop production. An epidemic is the culmination of all events affecting the initiation and development of a disease through the interaction of the genetic constitutions of host and pathogen with the critical environmental factors. The development of



epidemics requires: (1) abundant, viable inoculum of a virulent physiologic race of the pathogen, disseminated widely and rapidly at the proper times; (2) dense and extensive populations of a susceptible host in a receptive stage of development; and (3) optimal conditions of the environmental factors affecting production and germination of inoculum, penetration of the host by the pathogen, and rapidity of disease development. Generally, epidemics develop from successive cycles of dissemination, inoculation, infection, growth, and multiplication of the pathogen, as all requirements rarely are fulfilled simultaneously.

These interrelations were illustrated by the progression of an artificially induced epidemic of black stem rust in a field of Marquis wheat at St. Paul, Minn. in 1956 (see Fig. 23). The epidemic was initiated by inoculation with urediospores (summer spores) carried in oil. Inoculations of the wheat June 13 and 15 resulted in an initial incidence on June 23 of 15 to 20 pustules per stem, which was sufficient for abundant dissemination and inoculation of the wheat for favorable infection periods associated with heavy rain June 25, and lighter precipitation June 30 and July 1. Pustules from secondary infections were maturing by July 4 and an epidemic developed rapidly, resulting in completely diseased wheat within the following two weeks.

[J.B.R.]

#### DISSEMINATION OF PLANT PATHOGENS

Effective dissemination of pathogens requires that the inoculum be carried in viable condition to host plants that are susceptible to infection, and that the conditions at the time favor germination of the inoculum and infection of the host plants. Knowledge of the means of dissemination of a given

pathogen often is basic to control of the disease or diseases it causes.

**Dissemination by wind.** Spores are discharged into the air, or picked up by even the slightest of air currents, and carried by chance to susceptible hosts, sometimes only to nearby plants, sometimes for hundreds of miles.

**Local dissemination.** The conidia (spores) or sporangia of many of the downy mildew fungi are produced only at night or during periods of foggy or drizzly weather. They are very sensitive to drying and to sunlight, and usually can be effectively disseminated for only short distances. A single maize plant infected with *Sclerospora philippinensis* is estimated to produce from 1,000,000,000 to more than 5,000,000,000 spores of the fungus in a single night. Given a succession of nights favorable for production, spread, and infection, local epidemics of this and similar diseases can build up rapidly.

The fungus *Cronartium ribicola*, which causes white pine blister rust, can spread from wild currant plants (*Ribes*) to pines only by means of basidiospores which are small, delicate, and short lived. The basidiospores from wild currants can not be disseminated effectively more than 900 ft. If a stand of white pines is free from blister rust, it is necessary only to eliminate the *Ribes* bushes within the stand and up to 900 ft away to completely protect the pine from blister rust in the future. Many of the most valuable stands of white pines have been protected in this way.

The fungus *Ustilago tritici*, which causes loose smut of wheat, produces powdery masses of chlamydospores that replace the floral parts of infected plants. These spores are produced at the time the noninfected plants in the same field are producing their flowers. The wind carries the spores to flowers of adjacent plants and thus spreads infection locally. Although the spores can be carried long distances in viable condition, the short flowering period of the host plant ordinarily precludes effective long-distance dissemination. The same pattern is followed by a number of other loose smut fungi; and the use of smut-free seed, which eliminates inoculum from a given field or area, ordinarily is sufficient to protect the crops against these diseases.

**Long-distance dissemination.** East of the Rocky Mountains a more or less continuous belt of wheat is grown from northern Mexico to central Canada, and the wheat matures progressively from south to north. In many years, epidemics of stem rust of wheat, propagated by repeating urediospores of *Puccinia graminis* var. *tritici*, strike wheat in northern Mexico and southern Texas. The spores are carried northward by the wind, sometimes in short local advances, sometimes in hordes that infect wheat plants over an area of several hundred thousand square miles. Essentially the same course of events prevails with this disease in India and in portions of Russia.

**Dissemination by water.** A number of asphytic fungi are well adapted to spread by splash-

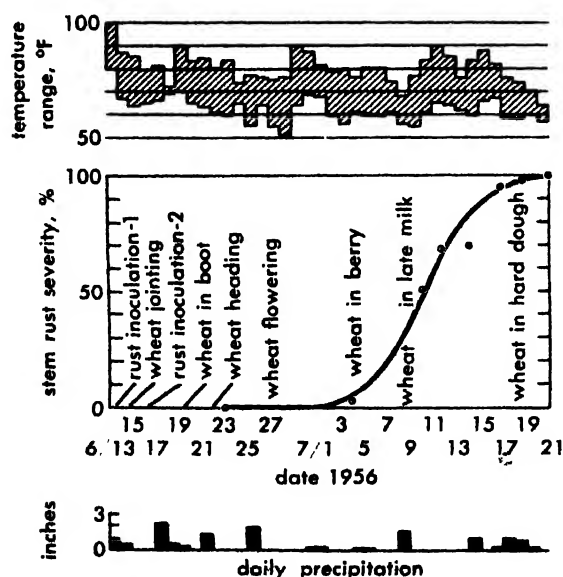


Fig. 23. Development of an epidemic of stem rust on Marquis wheat. (From J. B. Rowell, Oil inoculation of wheat with spores of *Puccinia graminis* var. *tritici*, *Phytopathology*, 47(11):689-690, 1957)

ing rain drops, but the dissemination of plant pathogenic fungi by this means has been little investigated. *Septoria* and *Phoma*, both of which cause a multitude of plant diseases, produce conidia in a sticky matrix within pycnidia (fruiting bodies). In wet weather the spores ooze out in tendrils. Spores of *Septoria acicola*, which causes a leaf disease of young loblolly pine, is known to be spread locally by splashing rain drops, and presumably most species of both *Septoria* and *Phoma* are so spread. They may be spread farther than locally, since spores of *Phoma* frequently are found in rather large numbers in the air. The dissemination of these and similar fungi by a combination of rain and wind deserves more study.

**Dissemination by insects.** Many of the most destructive virus diseases of economic plants are spread chiefly or only by insects, especially leaf hoppers and aphids (see HEMIPTERA). Effective spread usually is limited to the distance the insects can fly, but occasionally the insects, infected with the virus, may be carried many miles by the wind.

Bluestain of conifer trees and Dutch elm disease, both caused by species of *Ceratostomella*, are carried from infected to healthy trees by bark beetles. The beetles inoculate healthy trees with the fungus; the fungus invades and kills the trees, then the trees are invaded by greater numbers of the bark beetles which breed under the bark. Without the beetles the fungi could not be effectively disseminated, and without the fungus the beetles would do little damage, since ordinarily they cannot invade healthy trees in any numbers.

Insects are also responsible for development of epidemics of ergot in wild grasses and cultivated cereals. Primary infection of the flowers of these plants by *Claviceps purpurea* is by wind-borne ascospores (see ASCOMYCETES). This primary infection usually is very light, but within a few days the infected flowers produce large numbers of conidia in a sweet and sticky fluid. The fluid attracts certain small flies, which wallow in the spores, then visit other flowers and so spread the infection. Likewise pycniospores of many rust fungi are exuded in a matrix attractive to flies which carry the spores from one pycnium to another and insure fertilization of the rust fungus. Recently, grain-infesting weevils, moths, and mites have been found not only to carry inoculum of fungi that cause deterioration of stored grain, but also to increase the moisture content of the grain and thus enable the fungi to grow faster and cause more damage (Fig. 24)

**Dissemination by other animals.** *Endothia parasitica*, the fungus that causes chestnut blight, is carried by woodpeckers, and nearly a million spores of the fungus have been washed from the feet of a single bird. The woodpeckers probably not only spread the fungus locally, but also contributed, in their migrations, to the rapid north-south spread of this destructive disease. It is not known whether many birds in their local flights or long distance



Fig. 24. Fungi that cause deterioration of stored grain, growing from a larva of a grain infesting insect. The larva was surface disinfected to kill inoculum on the outside of its body. Several species of *Aspergillus* are growing out from the interior of the insect.

migrations regularly carry spores of plant pathogens and disseminate them effectively, but it seems highly probable. The role of birds, rodents, and other animals in the local and long-distance dissemination of plant pathogens requires more investigation.

**Dissemination by man.** Man has been responsible for rapid and widespread dissemination of some of the most destructive diseases of his economic plants. Before the communicable or contagious nature of many plant diseases was discovered about 1850, this spread was uninhibited and to some extent excusable. Even today, with rigorous inspection of plant materials in commerce and severe restrictions on the shipment of many plants and plant parts, dissemination of plant pathogens by man is by no means eliminated and probably will continue to be a large factor in world agriculture for many years to come. A few examples, grouped according to the causal agent of the disease, illustrate the nature of such spread, the damage that may be caused, and some of the difficulties involved in halting or reducing spread of plant diseases by man.

**Virus diseases.** Commercial growers of potatoes in the southern United States usually obtain their stock of "seed" or planting tubers from the northern states, chiefly because of the difficulty of producing virus-free seed stock in the South. But for many years the quality of seed potatoes grown in the North was erratic; some lots were heavily infected with one or more virus diseases that greatly reduced the yield of the crop grown from such stock. Since about 1920, seed certification departments have been established in the northern states that produce seed potatoes. Fields of potatoes are examined several times during the season by competent inspectors, and tubers are tested in the greenhouse during the winter to insure that no detectable virus is included in the seed stock. Some viruses, such as potato virus X, produce no symptoms in some varieties and may be carried along in these without detection. Such viruses, when spread

to other varieties, may cause considerable damage. All of the more than 50 known virus diseases of the common potato probably have been spread throughout the world by the shipment of diseased tubers. Similarly, virus diseases of stone and citrus fruits probably have been carried to all parts of the world where these crops are grown by shipment of diseased grafting stock. Sometimes these diseases do not produce any obvious symptoms until the trees are several years old. Special techniques are required to insure that grafting stock of material for clonal propagation such as cuttings is free of hidden viruses that may greatly reduce the productivity of the future crop.

**Bacterial diseases.** Ring rot of potatoes, caused by *Corynebacterium sepedonicum*, is spread chiefly by knives used to cut potatoes into pieces for planting and by contact of cut seed pieces or bruised whole potatoes with sacks, bins, or machinery contaminated with the bacteria. This disease was spread throughout the world by the shipment of infected potato tubers. A few infected tubers in a lot of many bushels may furnish sufficient inoculum to infect a large proportion of the seed pieces cut from these tubers, drastically reducing the yield of the crop grown from them. A somewhat similar bacterial disease of geraniums is spread chiefly by knives used to make cuttings for propagation, by the resultant contamination of soil, pots, and other equipment, and by the hands and clothing of workers. Constant strict inspection and sanitation are required to keep these diseases in check. Citrus canker, caused by *Xanthomonas citri*, was introduced from the Orient into Florida about 1910, apparently on diseased planting stock, and threatened to destroy the entire citrus industry there. It finally was eliminated by eradicating some 15,000,000 trees, a drastic, costly, but necessary measure.

**Fungus diseases.** Late blight of potatoes was carried from South America to and throughout Europe by the transport of diseased potato tubers, and later it was carried from Europe to North America. Both powdery and downy mildews of grapes, endemic on wild grapes in America, were carried by man to France on grape stocks imported to obtain new varieties. They quickly spread over Europe, for a time threatened the survival of the extensive grape and wine industries, and still add greatly to the cost of production.

White pine blister rust, native to Siberia, was carried into Europe on pines imported for growing in botanical gardens. About 1900 this disease was carried to the United States from Europe on infected seedling pines and now is found throughout the major white pine regions of North America. The Dutch elm disease was introduced into the United States from Europe about 1930. It has spread as far west as Wisconsin and has almost eliminated street plantings of elm in many eastern cities. A somewhat unusual case of plant disease being spread by man is that of a disfiguring and sometimes fatal canker of sycamore caused by a

species of *Ceratostomella*. This fungus was found to be spread almost entirely by means of spores in a wound dressing applied to the wood exposed when branches were cut off. The wound dressing was supposed to be fungicidal but actually served as an effective carrier of the inoculum of this fungus.

**Nematode diseases.** Although the evidence is largely circumstantial, it is likely that some of the most destructive plant parasitic nematodes have been imported into the United States in contaminated plant materials.

#### PLANT DISEASES IN STORAGE

Tubers, fruits, and fresh vegetables are subject to spoilage by a variety of pathogenic and nonpathogenic agents during storage and transit, and often this hazard remains acute up to the time of consumption. Seeds such as those of wheat, corn, barley, soybeans, and flax, which often are stored in bulk for months or years, also are subject to deterioration. At times, the losses in transit and storage equal those occurring while the plants are growing. In general, storage diseases are divided into those caused by nonpathogenic factors, and those caused by living organisms or pathogens.

**Nonpathogenic storage diseases.** Fruits and vegetables in storage suffer from a number of serious nonpathogenic or physiological diseases. Typically, these show up as discolored spots or areas on the surface of or within the affected parts, sometimes accompanied by collapse of the tissues, leaving pits on the surface or hollows within. These diseases are caused mainly by an excess of gases such as certain esters or carbon dioxide given off by the fruits or vegetables themselves, or by chemicals introduced into the storage rooms. These diseases may be controlled by maintaining proper storage conditions, including temperature, humidity, and aeration. Fruits, vegetables, and seeds harbor abundant microflora, and damage beginning from nonpathogenic causes may be increased greatly by subsequent invasion of the tissues by bacteria and fungi able to cause rapid decay.

**Pathogenic storage diseases.** Common fungi such as *Botrytis*, *Penicillium*, *Rhizopus*, and *Sclerotinia* invade and rot many fruits and vegetables. Losses up to 25% of a shipment between harvest and consumption are common in fruits such as oranges, apples, peaches, pears, and plums and in vegetables such as potatoes, sweet potatoes, tomatoes, and peppers. Bacteria, or a combination of fungi and bacteria, often rot stored potatoes and root vegetables. These diseases may be controlled by harvesting only sound, disease-free products, careful handling to prevent bruising, the use of clean containers, maintenance of low (about 40°F) temperatures in transit and storage, and at times the use of fungicides.

Grains stored in bulk are subject to invasion by a number of fungi, principally those in the genus *Aspergillus* (Fig. 25), which have the ability to grow at moisture contents in equilibrium with rel-

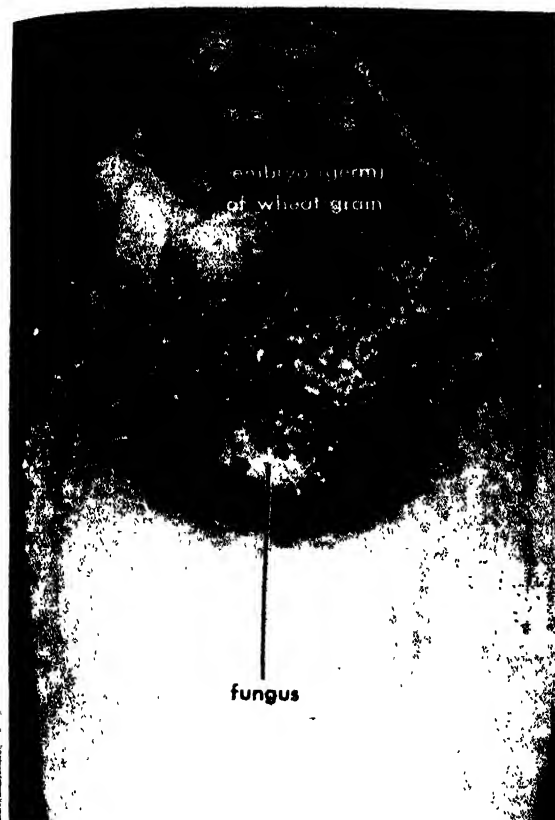


Fig. 25. A damaged grain of wheat, from a commercial storage bin, with *Aspergillus* (fungus) growing from the germ (embryo). Approximately 250,000 bushels of wheat in this bin were affected, with a loss of more than \$240,000.

relative humidities down to 70%. The resulting deterioration may not be detected until most of the damage has been done. Research is gradually making available, to men in the grain trade, facts and principles that will enable them to reduce such losses. [C.M.CH.]

**Bibliography:** F. C. Bawden, *Plant Viruses and Virus Diseases*, 3d ed., rev., 1950; C. M. Christensen, *The Molds and Man*, 1951; E. Gäumann, *The Fungi*, 1952; V. G. Lilly and H. L. Barnett, *Physiology of the Fungi*, 1951; E. C. Stakman and J. G. Harrar, *Principles of Plant Pathology*, 1957; J. C. Walker, *Plant Pathology*, 2d ed., 1957.

## Plant disease control

Control of plant diseases, by cultural methods, chemicals, resistant plant varieties, eradication of pathogens, and quarantines. For certain diseases these methods must be combined to give effective control.

**Cultural practices.** Cultural practices may control plant diseases by eliminating or reducing the effectiveness of the pathogen, or by altering the susceptibility of the host plant. This is done by tillage methods, use of fertilizers, crop rotation, and various sanitary practices designed to eliminate

sources of infection. See AGRICULTURAL SOIL AND CROP PRACTICES.

Some pathogens survive on crop residue, but are ineffective if buried by plowing, for example, the cereal head blight fungus on corn. Others, such as the cabbage black rot bacteria, live over winter, but will die out in a few years if a susceptible crop is not grown where they occur. Crop rotation, fertilization, and the growing of certain crops will encourage the establishment of organisms in the soil that destroy certain pathogens by antibiosis, for example, the cotton root rot organism which is reduced by liberal application of manure. See FERTILIZING; SOIL MICROORGANISMS.

Cultivation to eliminate weeds and variations in depth and time of planting also reduce losses from certain diseases. Sometimes unique tillage methods are employed, such as flooding fields with water to destroy the banana wilt organism. Excesses or deficiencies of both major and minor elements in fertilizers can cause disease or increase susceptibility of plants. See PLANT, MINERALS ESSENTIAL TO.

Sanitary practices, such as cleaning and disinfecting of farm equipment and destroying of crop residues harboring pathogens, often help to control specific diseases. [H.D.T.]

**Control with chemicals.** Chemicals are used to prevent or to reduce the severity of abnormal changes in plants induced by pathogenic organisms or by adverse environmental conditions.

**External plant protection.** Various chemical substances are applied to the surfaces of growing plants either to destroy superficially growing plant pathogens or to protect against infectious materials that might subsequently be deposited on the surface by wind, water, and insects (see INSECTA). The use of insecticides is included among external protectants because virus diseases of plants are commonly transmitted by insects. See INSECTICIDE; PLANT VIRUS.

External protectants may be applied to foliage, branches, stems, roots, seeds, corms, bulbs, tubers, and propagating stock in the form of sprays, dusts, dips, soaks, or slurries. Sprays are aqueous solutions, or occasionally oil suspensions, of fungicidal or bactericidal compounds applied with appropriate machines (see FUNGISTAT AND FUNGICIDE). Frequently surface-tension depressants and adhesives are added to liquid sprays to improve coverage and prolong adherence to plant surfaces (Fig. 1). Elaborate spray schedules have been developed to ensure adequate protection against such important diseases as apple scab, late blight of potatoes, leaf spot of bananas, and black spot of roses (see PLANT DISEASE). Dusts are dry formulations of fungicides, bactericides, and insecticides which are blown onto plants from ground equipment or from airplanes. The purpose of dusts is identical to that of sprays, but at times dusts are preferred for their ease of handling and in situations where weight and available water are important factors. Dusts are espe-



Fig. 1. Chemical control of potato late blight, Toluca Valley, Mexico. (Rockefeller Foundation)

cially important in the treatment of seed prior to planting or for the control of seedling blight and rot. Soaks and dips are solutions or suspensions of toxic compounds in which roots, corms, bulbs, and other vegetative plant parts are dipped for varying periods of time for disinfestation or for the control of specific diseases. Slurries are liquid seed treatments consisting of a milky suspension of chemical dusts in water.

**Internal plant protection.** This includes the correction of physiologic diseases arising from deficiencies of essential nutrient elements in crop soils and the use of systemic (internal) fungicides and bactericides. Deficiency symptoms indicate the lack of major or minor nutrients, such as nitrogen, phosphorus, potassium, sulfur, boron, iron, manganese, zinc, copper, and molybdenum. Deficiencies in major elements are readily corrected by the application of commercial fertilizers, and in microelements by the application of appropriate compounds to the soil, as foliage sprays, or more rarely, by direct injection into plants.

The concept of systemic plant protection is based on evidence that certain organic compounds are absorbed in minute quantities by crop plants, and either serve as protective substances against infection or stimulate living host cells to produce compounds which resist the attack of pathogenic microorganisms. It is hoped that certain of these materials may prove effective against the attack of fungi, bacteria, and perhaps viruses which are difficult to control by conventional methods. In fact, there is encouraging evidence that it may be possible to control stem rust of wheat, apple scab, and bean blight with systemic compounds.

The more common crop protectants include inorganic and organic compounds of sulfur, copper, chlorine, mercury, zinc, nickel, cobalt, and the quinones and phenols (see PHENOL; QUINONE). Those that appear to have systemic effects include salts of heavy metals such as iron, mercury, zinc, and lead, the sulfamates, salicylic and picric acids, dithiocarbamates, certain fumigants such as methyl bromide, and the herbicides 2,4-D and 2,4,5-T (see FUMIGANT; HERBICIDE). A number of antibiotic substances extracted from microorganisms have been shown to act as systemic protectants (see ANTIBIOTIC). Examples are Actidione, streptomycin, and antimycin from *Streptomyces* species, fun-

gocin from a species of *Bacillus*, and viridin from the fungus *Trichoderma* (Fig. 2).

**Weed control.** Because weeds are plant pests to the extent that they compete with crop plants and serve as reservoirs of plant pathogens and hosts to insect vectors (carriers), the plant pathologist includes weed control in the general area of plant protection. For the chemical control of weeds complex compounds known in the trade by such names as 2,4-D, 2,4,5-T, TCA, IPC, and 2,4-DB are used. Some herbicides are toxic to the weeds treated whereas the selective herbicides are principally growth-promoting substances which overstimulate living tissues and ultimately cause their death.

[J.C.H.]

**Disease resistance.** Growing resistant varieties of crop plants is the best and most economical way of combating plant diseases. No capital, labor, or time need be expended in applying protective fungicides if the crop varieties grown can resist the ravages of disease.

Varieties that have no natural resistance to a pathogen often escape disease infection and damage merely because they mature early or because of their growth habit. For example, certain wheat usually escape rust because they ripen early. Bush-type beans are less frequently attacked by the white mold *Sclerotinia* than are the vine types.

Better understanding of the specificity of resistance comes as more is learned about parasitism and disease processes, physical factors and mechanisms, chemical components and combinations, and enzyme systems and energy exchanges. Even without full understanding, much has been done in utilizing and combining the kinds of resistance, studying the inheritance of resistance, investigating its variability, and practical testing for its dependability.

**Heritable resistance.** A true natural resistance, however, is a heritable character, sometimes governed by a single gene, sometimes by many genes. Resistance is based on structure, protoplasmic properties, metabolic activities, and physiological functioning of a variety; it may vary qualitatively with external environmental factors. Usually it is effective against a specific pathogen, but not against all pathogens. Immunity, connoting com-



Fig. 2. Control of downy mildew of lima beans with a streptomycin sulfate compound, Agriotrop. (a) Untreated, 100% infection. (b) Sprayed with 100 ppm of streptomycin, excellent control. (W. J. Zimmerman, USDA)

plete freedom from a pathogen and a disease, represents absolute resistance. In most plants, however, there are varying degrees of resistance rather than immunity. Some varieties of wheat and potatoes have a combination of characters that protect them fairly well under many conditions in the field against stem rust and late blight, respectively. This generalized kind of resistance, not yet completely understood, is called adult-plant resistance, mature-plant resistance, or field resistance. Hypersensitivity also has the practical effectiveness of a high degree of resistance, although it actually is a supersusceptibility which results in the sudden death of the first few cells attacked but establishes conditions in which the remaining host cells are protected from continuing attack.

Genes responsible for resistance have been determined and utilized in some cases, and the inheritance of the resistant character has been studied. A single dominant gene controls type-A resistance to yellows in cabbage. At least four dominant or major genes, all derived from *Solanum demissum*, control late blight resistance in potato, and there is some evidence that one or two other major genes exist. Many other genes (modifying or perhaps minor genes) are involved in the field resistance to potato late blight, whereby a variety is attacked by the pathogen but produces an acceptable crop because the damage from the attack is seldom severe. There are genes in flax that govern resistance to the various races of flax rust and, correspondingly, there are genes that govern virulence in the races of the flax rust fungus.

*Resistance attributes.* Some investigators seek the reasons for a variety's resistance only in the incompatibilities of the two associated metabolic systems, that of the host variety and that of the living pathogen. Other workers regard as contributors to resistance all of a variety's characters or attributes that help to ward off or retard entry of a pathogen, restrict or inhibit a pathogen's progress within the variety, or actually destroy the invader.

Certain resistant varieties and species of plants have a single attribute that thwarts a particular pathogen in its attack; others have many attributes playing roles of varying importance as they contribute to the variety's resistance to a specific pathogen. A thick, tough epidermis may prevent direct penetration. Prolonged closure of stomata or extremely small stomata may bar the way for some pathogens that normally enter these natural openings of a plant. Occasionally a stomatal configuration that interferes with formation, retention, and continuity of water films excludes a pathogen. Lenticels with compact cell organization and with capacity for rapid suberization of cells retard the invader. Cell walls with much crude fiber in the secondary thickenings often impede the advance of fungus hyphae; and lignified cell walls may be impassable barriers for some pathogens confined to the rigid plant tissues. Water-soluble chemicals diffusing from some plant parts into the sur-

rounding soil or water are toxic to some pathogens and inhibit their growth. Alkaloids toxic to certain root-rot organisms are components of some resistant varieties, and strong acids and volatile sulfides are the protective chemicals in others. Phenolic compounds, such as tannins, catechol, and chlorogenic acid, are often more abundant in resistant varieties than in susceptible varieties, and although some phenolic substances may be bound in the host cells in nonlethal form, there are other forms of these compounds which are actively toxic to pathogens.

*Variations in resistance.* A simple and general explanation of resistance is not possible. A variety with resistance to one kind of pathogen is not necessarily resistant to other pathogens that usually attack that plant. Different varieties and species of plants may suffer from one disease or from several diseases; and thousands of pathogens, particularly those among the fungi, bacteria, viruses, and nematodes, cause various kinds of disease in plants (see NEMATODA). The reaction of a variety to one disease may be independent of its reaction to several other diseases. For example, the wheat variety Thatcher was valued for its stem-rust resistance in the 1930s, but it lacks adequate resistance to leaf rust, scab, and root rot.

A resistant variety may not be resistant to all representatives of one pathogen. Just as a single species of a crop plant often comprises many agronomic or horticultural varieties, so may a single species of a pathogen comprise several varieties and almost innumerable strains or races that differ genetically, physiologically, and in pathogenic capabilities even though they look sufficiently alike to be recognized and classified as the single species. This physiologic specialization exists in several pathogens, such as many of the cereal rusts, the late-blight *Phytophthora*, the wilt *Fusaria*, the root-rotting *Rhizoctonia*, the bean mosaic, the tobacco mosaic, and the sugarbeet curlytop viruses. This phenomenon probably is of more general occurrence than has been demonstrated, and wherever it exists it complicates crop improvement and the breeding of new disease-resistant varieties. In many instances the resistance of any variety must be considered in relation to an individual race or strain of a specific pathogen.

*Environmental factors.* External factors frequently change the quality of a disease reaction. Temperature, light, nutrients (both major and minor elements), carbon dioxide concentration, and hydrogen ion concentration (acidity) may effect changes. Sometimes the changes are slight, sometimes extreme. A notable example of the latter is common in some of the wheats from Kenya and in some of the varieties of oats. These cereals may be highly resistant to individual races of stem rust at one temperature and susceptible to the same races if the temperature is only 15°F higher. Thus, the resistance of a variety frequently must be considered in relation to a certain race of a pathogen and also in relation to a certain environment.



**Wild and cultivated plants.** Wild plant species that have undergone natural selection frequently have more disease resistance than do cultivated plants, especially if there has been long association of the plants with the disease organisms. *Solanum demissum*, a wild potato with small tubers, is more resistant to late blight disease than are the cultivated varieties of *S. tuberosum*; and many other species of *Solanum* growing throughout the highlands of South America are being investigated as sources of resistance to late blight or some other diseases. The small-fruited wild tomato, *Lycopersicon pimpinellifolium*, has more resistance to a number of diseases, such as bacterial wilt, fusarium wilt, verticillium wilt, bacterial canker, gray leaf spot, leaf mold, and the spotted wilt viruses, than has the cultivated tomato *L. esculentum*. Sometimes, by selection and by hybridization, the disease resistance of a wild plant can be incorporated into varieties suitable for cultivation.

Cultivated crops sometimes are heterogeneous populations from which resistant individuals can be selected and propagated. The simple selection of survivors in a plant population subjected to repeated severe attacks of such soil-infesting pathogens as *Fusarium* or *Verticillium* has been one means of obtaining varieties of cabbage, cotton, flax, melon, and tomato which are resistant to various wilt diseases. Growing flax year after year in a soil plot infested with all available and known strains of the fusarial flax-wilt fungus, propagating from the survivors, and again growing the progeny in the infested soil eliminates the wilt-susceptible portions of the populations. By persistent selection of this kind, flax wilt has been controlled in the United States for about 50 years.

By hybridization many resistant varieties have been produced in wheat, oats, barley, cotton, flax, potato, tomato, bean, cabbage, melon, cucumber, sugar cane, sugar beet, tobacco, and various other crops. Simple and complex crosses, polycrosses, species crosses, and backcrosses have been made, and isogenic lines have been produced in some instances. If genetic characters for disease resistance are available, they can often be combined with many other desirable characters in spite of difficulties encountered because of lack of synchronized flowering in the materials, pollen sterility, and linkages. The search for resistant materials may cover wide territory. For example, sugar canes from Java and India supplied the mosaic resistance that was needed for canes in the United States, and wheats from Australia and Kenya furnished resistance to race 15B of stem rust for many new wheat hybrids in the United States, Canada, and Mexico. New sources of resistance are continually being sought in the varieties and genic lines assembled in world collections of various crop plants.

In times past when some disease has threatened destruction of certain crops and discontinuance of their cultivation, resistant varieties have made possible their continued production. Today the aim is greater foresight by planning adequate testing of

resistant varieties, by having various kinds of resistance at hand and available, and by facilitating rapid replacements or recombinations of crop materials.

[H.H.A.]

**Eradication campaigns.** These are designed either to eliminate recently introduced pathogens completely, or to protect economic plants by destroying alternate, or weed, hosts. Success in eliminating pathogens depends on early detection of the pathogen and on the efficiency of eradication measures.

Attempts made in the United States to eradicate chestnut blight and the Dutch elm disease were unsuccessful. The citrus canker disease, however, was eliminated from Florida by burning infected trees. Flag smut of wheat, which was introduced locally into Mexico, was also successfully burned out. Similarly, persistent eradication of infected plants has helped restrict many diseases. See FRUIT (TREE) DISEASES.

Certain rusts can be controlled wholly or partly by eradicating alternate hosts. For example, the destruction of red cedars near apple orchards protects apples against the *Gymnosporangium* rust because this rust cannot maintain itself on either host alone. To help control stem rust of wheat and other small grains, the growing of barberries, *Berberis* spp., has been prohibited by law in some countries. Denmark began a successful campaign against barberries in 1904 and in the United States about 500,000,000 barberries have been destroyed since 1918, with substantial reduction of the stem-rust menace (Figs. 3 and 4). Likewise in the United States white pines and other susceptible species are partly protected from blister rust by eradicating nearby currants and gooseberries, *Ribes* spp.

Like legal public health measures for human beings and for domestic animals, those for plants are essential in keeping many diseases in check.

[E.C.SN.]

**Quarantines.** Plant disease quarantines are legal measures taken by Federal or state governments to prevent the introduction of foreign plant diseases or pests into an area. Quarantines are based on the philosophy that government has the right and ob-



Fig. 3. Hormone-type chemical sprays are used for killing rust-susceptible barberry plants. This process is much faster and less costly than common salt. (USDA)

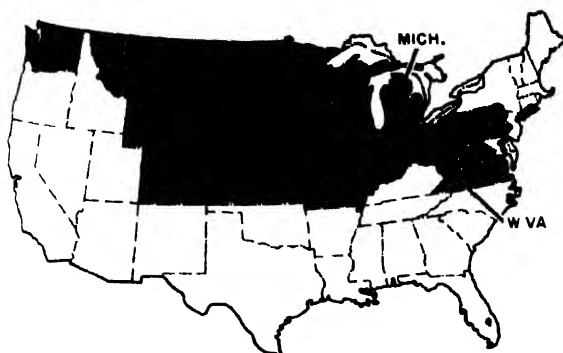


Fig 4 More than 513,500,000 rust-susceptible barberry bushes have been destroyed on 153,000 different rural and urban properties in the 19 barberry-eradication states. (USDA)

ligation to protect its agricultural resources and industry from the destructive effects of exotic plant diseases and pests.

The transportation of plants and plant parts was long a matter of private concern, with the result that many plant pathogens became widely distributed by international travelers or through unrestricted trade channels. The dangers of this situation became dramatically apparent following the accidental importation of the chestnut-blight fungus into the United States from Asia between 1900 and 1905 and the ultimate destruction of the American chestnut forests.

As a consequence of this and other bitter lessons, the United States Government in 1912 passed the national Plant Quarantine Act. Today essentially

Nations have enacted protective quarantine regulations. Quarantine laws authorize Federal or state officials to intercept and inspect shipments of plant materials and to release, fumigate, or confiscate the shipment in accordance with legal provisions. Quarantine inspectors are stationed at ports of entry, border stations, and at receiving and distributing points for freight and mail.

**Value of quarantines.** The value of quarantines has long been disputed. Antagonists claim that man is unable to prevent the movement of microscopic pathogens, that many quarantines are scientifically unsound, and that on occasions quarantines have been used as economic sanctions in restraint of free trade and have caused unnecessary economic losses. Supporters insist that even though not 100% effective, quarantines do prevent the introduction of many pests and diseases and retard the movement of others, giving scientists time to combat them before they become well established, that quarantines annually save the agricultural industry millions of dollars, and that these economic gains are many times greater than any possible business losses resulting from the application of quarantine measures.

Improvements in the practice of quarantining may provide assurance that all quarantines will be established on sound biological bases for maximum effectiveness, that they can be lifted with equal fa-

cility when it becomes clear that they are no longer necessary, and that, insofar as possible, new quarantine laws would be preceded by international consultation in an attempt to obtain mutual agreement and to ensure minimum disruption of the international exchange of commodities.

**International disease protection.** Joint efforts are made by nations to protect their agricultural resources and industries without impairment of exchange of commodities. Ideally, available knowledge on plant pests and pathogens is utilized to devise methods to limit their geographic spread and to prevent the outbreak of epidemics. Changes in cropping patterns, in trade agreements, and in the distribution of pests and pathogens necessitate a continuing program consisting of (1) annual plant disease surveys by the several nations with free exchange of results; (2) the rigorous practice of local sanitation and plant protection; (3) the prompt distribution of resistant varieties of crop plants; (4) the exchange of information in improved control measures; and (5) international consultation with respect to the establishment and enforcement of quarantines.

International plant protection can be successful only when regulatory activities are fortified by scientists investigating the etiology of plant diseases, life cycles of pathogens, host-parasite relationships, and chemical and other control measures. Exchange visits by scientific personnel further strengthen understanding and lead to logical and amicable agreements. Among international organizations active in plant protection are the Food and Agriculture Organization of the United Nations, and the International Commission on Plant Disease Losses.

[J.C.H.]

**Bibliography.** E. C. Stakman and J. G. Harrar, *Principles of Plant Pathology*, 1957; A. Stefferud (ed.), *Plant Diseases*, USDA Yearbook Agr., 1953; J. C. Walker, *Plant Pathology*, 2d ed., 1957.

## Plant evolution

That phase of evolution dealing with the origin and development of the plant kingdom. This article discusses evolutionary processes or dynamics, and evolutionary history or phylogeny.

**Evolutionary processes.** The processes of plant evolution—mutation, genetic recombination, natural selection, and reproductive isolation—are the same as in all forms of life, but profound differences between animals and plants in their mode of life, basic structure, and individual development bring about corresponding modifications in the mode of action and relative importance of these four processes. Since plants either manufacture food by photosynthesis, or, as saprophytes or parasites, absorb it after excreting digestive enzymes into the surrounding medium, they lack integrated systems for ingestion, digestion, and excretion. Being essentially nonmotile, they also lack the stimulus-response systems which in animals show the highest type of integration. Because they absorb rather than ingest, plants require an increase of ex-

ternal surface area as their total volume increases, whereas animals function best if their body is compact, and their increase in volume is accompanied by a corresponding increase in surface area of internal organs. Hence, animals require a carefully integrated and centralized circulatory system, while plants need only a relatively loosely organized, decentralized system of tissues for conducting food and water.

The compact body of animals requires an integrated pattern of development, known as the closed system of growth, in which all tissues and organs pass almost simultaneously through stages of embryonal primordia, youth, maturity, and old age. The less-integrated organization of plants is produced by the open system of growth, in which the leaves, branches, and flowers are produced serially by specialized embryonic tissues or meristems, so that embryonic, youthful, mature, and senescent stages may exist simultaneously in different organs of the same individual, and the same type of organ may be produced repeatedly over an indefinite period of time. The developmental cycle of any particular organ in plants is shorter, less complex, and less well integrated with other organs than in animals. Consequently, plant organs can be modified much more profoundly by both environmental and genetic effects than can animal structures without producing inviability.

Because of their general lack of motility, plants have evolved special devices for the dissemination of propagules, such as spores and seeds (see POPULATION DISPERSAL). In addition, the flowering plants have evolved elaborate structures which promote cross pollination through the aid of animals. Together with the structures which protect the spores and seeds, the progressive elaboration of these methods of cross pollination and dispersal of spores and seeds forms the principal thread around which the evolution of land plants is centered. A secondary thread is formed by the vascular system, which serves for conduction and support. On the other hand, the modifications of the outward form of leaves and stems, as well as of the inner physiological condition of their cells, which adapt plants to different habitats and modes of life, are of such a general nature that they are repeated in an almost parallel fashion many times in unrelated groups of plants. They are consequently much less reliable as signposts for the pathways of evolution than are corresponding maintenance structures in animals.

In addition, the dioecious condition (that is, individuals distinct as to sex—male or female) is far less widespread in plants than in animals. In some plants such as mosses monoecious types (having individuals both male and female) have been derived secondarily from dioecious ancestors.

All of the above-mentioned characteristics of plants have facilitated evolution through recombination of genetic characteristics derived from different adaptive systems. Hybridization between subspecies and species is particularly common in plants, and the hybrid derivatives are often vigor-

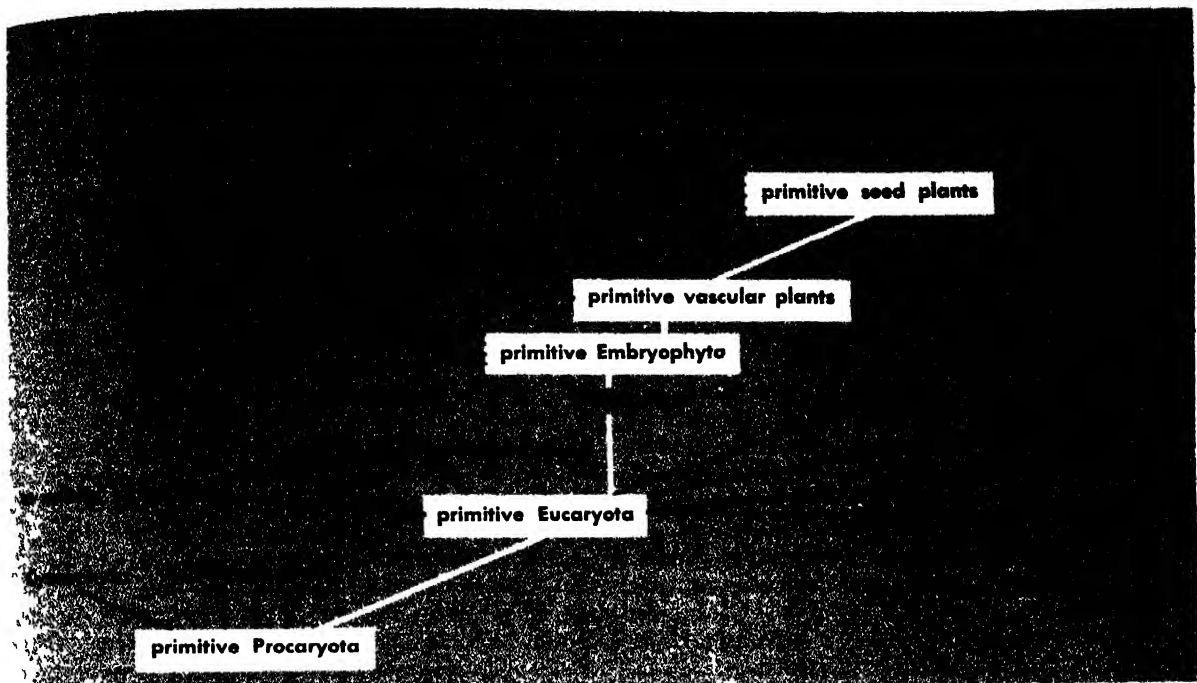
ous and well adapted to new habitats. They can become stabilized in three different ways; by segregation of relatively homozygous intermediate types, by introgression, and by polyploidy. The majority of species of higher plants have probably evolved at least in part by one of these processes.

**Evolutionary history.** The evolutionary history of the plant kingdom is outlined in the diagram (see illustration). Five levels of advancement are recognized, each based upon an adaptive system of offering new opportunities for adaptive radiation. Each level was first reached by a group with generalized or primitive characteristics. Through adaptive radiation, this generalized group gave rise in each case to a number of evolutionary lines or phyla, the modern representatives of which are very different from each other.

The lowest stage is the Procaryota, or organisms with a primitive type of nucleus lacking a clearly defined membrane; their genes are organized into aggregates less sharply defined and less highly integrated than the chromosomes of higher organisms; their nuclear division is less complex than typical mitosis. Sexual reproduction, when present involves a minimum of morphologically specialized structures. Differentiation is confined to the formation of resting cells or spores. The primitive Procaryota gave rise to the Cyanophyta, or blue-green algae, and to various lines of the Schizomycophyta or bacteria (see SCHIZOMYCETES). Fossil blue green algae have been found in rocks 1,000,000,000 years old.

The primitive Eucaryota, or organisms with a well-defined nuclear membrane, chromosomes, and mitotic cell division, were unicellular. Their modern counterparts are flagellate organisms which are classified either as algae or as protozoa. They gave rise to eight different phyla of plants, most of them primarily aquatic, which differ from each other chiefly in their methods of synthesis or absorption of food. The autotrophic phyla, which collectively are termed algae, have developed different types of pigments associated with their chlorophyll. These pigments enable them to absorb and use for photosynthesis light of wavelengths characteristic for each phylum. The heterotrophic phyla, which are saprophytic or parasitic, are the Myxomycota and the Eumycota, or fungi. Certain Phaeophyta, or brown algae, and the most advanced basidiomycetes or mushrooms among the fungi, have evolved well-differentiated tissues, though these are less elaborate than are those of most Embryophyta. The most common life cycle in primitive Eucaryota is the haploid type, in which meiosis immediately follows fertilization and zygote formation. Strictly diploid life cycles are best known in the diatoms (Chrysophyta) and the Fucales (Phaeophyta). Certain Rhodophyta, Phaeophyta, and Chlorophyta possess an alternation of two morphologically similar generations, a haploid gametophyte and a diploid sporophyte.

The next stage of advancement, reached by certain derivatives of the Chlorophyta, involves elaboration of the archegonium, a structure which pro-



Phylogenetic diagram of the plant kingdom.

nects and nourishes the zygote and young embryo. Because it frees one generation of the life cycle from dependence on water for early development,

evolution was the most important single step taken by higher plants in conquering the land. It is assumed that these earliest Embryophyta were different from both of their modern descendants, the Bryophyta and the Tracheophyta, partly because there are such great morphological gaps between these modern groups, as well as between them and living Chlorophyta; and partly because of recently discovered spores abundant in strata of middle and early Paleozoic age, and which apparently belonged to land plants. They testify to the existence of an extensive early land flora which has left few or no fossil remains. The bulk of recent evidence favors the homologous theory of the origin of the alternation of generations, according to which these earliest Embryophyta, like many Chlorophyta, possessed two morphologically similar generations. From them, evolution proceeded in one direction toward the Bryophyta, with highly differentiated gametophytes and reduced, parasitic sporophytes.

In the opposite direction there evolved the Tracheophyta, with reduced but free-living gametophytes and with enlarged sporophytes possessing well-differentiated vascular systems. These constituted a new level of advancement, because their tissues for support and conduction permitted them to attain large size as land plants, and to dominate their environment. During the later Devonian and the early Carboniferous periods they gave rise to many adaptively radiating lines, which included the first trees. Although the earliest dominant Tracheophyta were spore-bearing, being related to the modern club mosses (Lycopsidea) and horsetails

(Sphenopsida), seed-bearing plants already existed in smaller numbers in the Devonian period. The ferns also began their evolution at that time, and ever since then have been abundant in certain habitats.

The earliest seed plants and their descendants constitute the fifth and highest level of advancement achieved by the plant kingdom in generalized adaptive characteristics. They evolved highly efficient structures for protecting the developing gametophyte and nourishing it from the maternal sporophyte, for cross fertilization following aerial transport of relatively drought- and temperature-resistant male gametophytes (pollen grains), and for protecting and nourishing the embryo by the maternal sporophyte. During the moist, equable climate of the Carboniferous period, spore-bearing and seed-bearing plants were almost equally abundant, but the advent in the Permian period of cold and arid climates gave the more resistant seed plants a tremendous advantage. They evolved along lines leading to several extinct groups, to the modern orders of gymnosperms (Cycadales, Ginkgoales, Gnetales, Coniferales), and to the angiosperms, or flowering plants.

The origin of angiosperms is unknown. Their first undoubted fossils are from early Cretaceous strata, 120,000,000 years old. After this time they became exceedingly abundant. Older fragmentary remains, mostly of wood and pollen, but including a few leaves, have been identified as angiosperms by various paleobotanists. The earliest are from the Triassic period, and suggest that the angiosperms may be 175,000,000–200,000,000 years old, and that their unquestioned fossil record comprises only the latter two-thirds of their evolutionary history. During the first 60,000,000–80,000,000

years of their existence, angiosperms may have grown on mountain tops or hill slopes where preservation as fossils is difficult or impossible. See PALEOBOTANY.

The ancestors of the angiosperms may have belonged to the extinct pteridosperms, or seed ferns. Anatomical evidence suggests that the most primitive angiosperms were trees or shrubs, but evidence from both fossil and the morphology of living forms indicates that herbs related to the waterlilies (Nymphaeaceae) and buttercups (Ranunculaceae, Paeoniaceae, *Kingdonia*) appeared very early in angiosperm evolution. Dicotyledons are in many respects more primitive than monocotyledons, but evidence such as the occurrence in Triassic strata of fossil leaves which may be palms indicates that monocotyledons diverged from the primitive angiosperm stock at the beginning of its evolution, and evolved simultaneously with dicotyledons. See BRYOPHYTA; CHLOROPHYTA; CYANOPHYTA; EMBRYOPHYTA; EVOLUTION, ORGANIC; GYMNOSPERMAE; LIFE, ORIGIN OF; PLANT KINGDOM; POLYPLOIDY; PROTOZOA; PTEROPSIDA; SPECIATION; THALLOPHYTA; TRACHEOPHYTA. [G.L.S.]

**Bibliography:** A. J. Eames, *Morphology of Vascular Plants*, 1936; G. M. Smith, *Cryptogamic Botany*, vols. 1 and 2, 1955; G. L. Stebbins, *Variation and Evolution in Plants*, 1950; A. L. Takhtajian, *Origins of Angiospermous Plants*, 1958.

## Plant facilities

The physical properties owned and used by industry constitute the plant facilities. These include land and site improvements such as roads, rail extensions, parking lots, and fencing; buildings such as factory, warehouse, office areas; other structures such as docks, liquid-storage tanks, incinerators, and gas generating systems; and machinery and equipment used to produce or condition the products of the plant or to support the producing (or conditioning) processes.

**Location.** An industrial plant is often located merely by preference of one or a few owners of the business. Large companies make a detailed study before deciding on locations for new plants. The decision is based on many factors; most of them are economic and relate to the costs of transportation (for incoming materials and outgoing product), of materials, of labor, of taxes. Many industries depend directly upon their source of raw materials and the market area for their products. Other industries consider first the availability of satisfactory features like space, freedom from adjacent dirt and fumes, trained labor supply, water and utilities. Another factor is community climate or the attitude of the people and their leaders.

A thorough plant-location study includes evaluating the availability, suitability, and long-range cost of landsite, raw materials, power, fuel, water, utilities, market area, transportation facilities, labor supply, community conditions, taxes, public relations value, laws or regulations, and external hazards.

**Plant layout.** The layout of an industrial plant embraces the arrangements and orientation of the physical facilities, including storage features and supporting services, used by the company in the production of its products. Planning a new or rearranged layout involves four basic phases.

1. Determining the location of the area to be laid out is not necessarily a plant-location problem. More often it is one of analyzing whether the new department or expansion should go on the north side, on the third floor, or in a separate building.

2. Establishing the general over-all layout involves relating the major activities or departments to each other and allocating the necessary area to each. At this phase, major features of the producing machinery, materials handling equipment, utilities, and the building itself should be incorporated.

3. Planning the detailed layout includes the locating and orienting of each machine, working operator, material-in-process container, and supporting service.

4. Installing the layout plan involves coaching or training personnel in the procedures and methods on which the proper functioning of the layout has been planned, and moving, placing, and hooking up the machinery and equipment.

These four phases fall in chronological sequence but always overlap each other. Also, these four phases make logical check points or organizational divisions for the supervisor of plant layout work.

Effective layouts are based on flow of material to allow sequential movement of the material being produced or conditioned. As a result, determining the flow is the heart of most layout projects. As the number and diversity of parts or products increase, the complexity of flow analysis grows and the methods of analysis change.

Diagramming the flow and then assigning to the diagram (or diagrams) the space required by the activities are steps that follow determining the flow. These steps lead to a physical arrangement of space. Integration of supporting services with the flow pattern and modification of the arrangement in accordance with practical limitations of the handling and storage facilities, machine utilization, building features, flexibility, opportunity for expansion, and personnel needs and conveniences lead to the desired layout. Alternative plans are usually developed. By comparative evaluation the most effective is selected.

Layouts are most readily visualized by making a three-dimensional model of the proposed plan. But this is not always necessary. Two-dimensional scale templates representing the space—and the individual pieces of machinery and equipment in detail layouts—are always practical. With modern plastic materials, many different layout plans can be made and reproduced with a minimum of draftsmanship and drawing time.

Classical plant layouts center about either the product or process, when forming or treating materials is involved. Layout by product all facilities required for one product out by process locates all similar p:



ment together. The former is better for high volume work; the latter for high variety of work.

In assembly, the choice is between moving the major component progressively to points where other parts are assembled to it—line production—or fixing the location of the major component and bringing other parts to it. Here again volume and variety essentially determine the choice.

**In-plant transportation.** Materials handling and plant layout go hand in hand. One can seldom plan or change one without affecting the other.

Materials handling is a universal production problem. To form, condition, or assemble the materials, they must be moved to and from the point of operation.

Planning effective materials handling involves the selection of a basic handling system. Individual handling equipment and containers are then fitted into this system. The particular equipment may be hand trucks, overhead cranes, or conveyors. An integrated materials-handling operation depends on the containers and attachments for the handling equipment (clamps, hooks, brackets, and the like) being planned for interchangeable use.

Although improvements in equipment are constantly being adopted, the basic handling system of a company should remain relatively stable. Industries with high initial investment in facilities and with relatively fixed products, processes, and equipment establish a system of handling around which the plant is constructed. Factories with frequent model changes remain flexible and are constantly examining their handling methods for improvement (see MATERIALS HANDLING).

**Production lines.** A production line is an arrangement of work places in the sequence of operations. In its optimum form, a production line moves the material through a series of balanced operations smoothly and continuously at a uniform rate of flow with each operation being located immediately adjacent to the ones which precede and follow it.

The big savings that come from using line production include reduced material handling, ease of production, control and supervision, improved work-area methods, ease of training workers, reduced inventory in process, and shorter production time. On the other hand, production lines require a substantial volume of a reasonably standardized or at least similar—product. As a result, the marketing and product design of a company greatly influence the possibility and nature of its production line. Other limitations include delays due to breakdowns or interruptions; idle workers due to unbalanced operation times; and greater investment in machinery because its utilization is seldom as great as in a layout by process.

In some cases, the physical ability and economic practicality to move the product from one operation to the next limits the application of a production line.

For these reasons, production lines for assembly are easier to set up than those for forming or fabrication. There are many variations of the line con-

cept, with the ultimate seldom proving practical. Flexibility of the line to accommodate changing products, materials, processes, schedules, and personnel is far too important to permit very many companies to go all the way to the completely synchronized and automatic production line.

A production line is scheduled and operated as a unit, not as individual operations. Its workers generally need be trained to do but a single operation. These two facts can cause considerable change in the type of employee and the procedures of operating an organization when it converts its operations to production lines.

**Maintenance.** Vital to any low-cost industrial operation is good maintenance of its facilities. Particularly is this so as industry continues to substitute power operations for hand labor. The breakdown of but one key operation can sometimes cause the shutdown of an entire plant because of its synchronization with subsequent operations.

Preventive maintenance such as cleaning, adjusting, exchanging, and lubricating on a programmed basis eliminates or substantially reduces shutdowns due to machine failure. This is perhaps the most important function of a plant's maintenance group. Other functions may include inspection, repair, overhaul, reconstruction, salvage, waste disposal, plant protection, and storekeeping.

So diverse are the problems of maintenance that one person must be assigned its responsibility. Maintenance should be planned, scheduled, and efficiently executed, but its costs must be subject in part to the availability of funds. Its accomplishment should be measured by performance rather than the amount spent, for frequently not enough money is spent on maintenance of machinery, equipment, and plant for efficient over-all operations.

**Safety and fire prevention.** The cost to industry each year of interruptions due to accidents and fires is substantial, purely aside from non-economic considerations.

Insurance is one way of minimizing losses; fire extinguishers and first-aid stations are another way. However, elimination of the causes and opportunities for an accident to occur is a more direct and rewarding approach.

Safety rules—established for the specific plant involved and consistently enforced—are well worth the effort. Detailed rules for specific equipment, located where they are read before the machine is used, are another precaution.

For effective fire control, every plant needs a fire-detection and alarm system, sprinklers, mains, hydrants, and extinguishers. Specific equipment depends on the nature of the industry's products, processes and plant. See FACTORY; INDUSTRIAL ENGINEERING; PRODUCTION METHODS. [R.M.]

## Plant fermentation

A form of plant metabolism in which carbohydrates are partially degraded without the consumption of molecular oxygen. Approximately 8.5% of the potential energy of the carbohydrates is released;



the remainder of the energy is transferred to the organic compounds resulting from the fermentation. Of the energy released, approximately 44% is in forms available for cellular work, and the remainder is degraded to heat. See CARBOHYDRATE METABOLISM.

**Products of fermentation.** The products of fermentation vary with the organism, but in higher plants the usual products are ethyl alcohol and carbon dioxide. In some cases small amounts of lactic acid are produced, and other organic compounds of uncertain structure may be formed, for example, in potato tubers.

Certain incomplete oxidations, such as the oxidation of ethyl alcohol to acetic acid by *Acetobacter pasteurianus* or of sucrose to citric acid by *Aspergillus niger*, are called fermentations in industry. However, because these processes are strictly dependent upon molecular oxygen, they are not fermentations. See INDUSTRIAL MICROBIOLOGY.

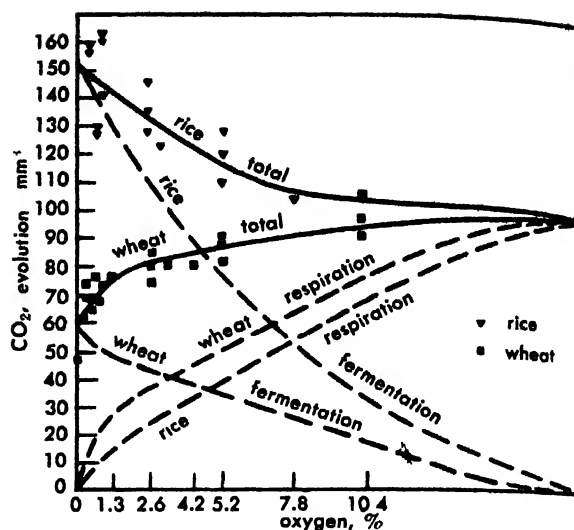
Fermentative capacity is widespread and is known in many lower plants, including yeasts, certain fungi, and algae, and in most higher plants and animals. In higher plants, fermentation often occurs at high velocity in embryonic structures, such as seeds, seedlings, young roots, and shoots. The velocity often decreases during development, and may cease entirely in mature stems and leaves of certain plants. The occurrence and rate of fermentation are illustrated in the table.

**Effect of oxygen on fermentation.** The partial pressure of oxygen has a marked effect on fermentation. Except in a few obligate anaerobic bacteria, fermentation ceases when the oxygen pressure in the cells exceeds values of 3–10 mm Hg. The suppression of fermentation by oxygen is known as the Pasteur effect. The external pressure of oxygen that just inhibits fermentation, known as the extinction point, is as illustrated. This pressure is considerably higher than the internal pressure, and is often 20–40 mm Hg in range. When rice or other seeds germinate under water, the tissues may lack oxygen, and fermentation results. If the intercellular spaces become injected with water, the cells often become deficient in oxygen, with a resulting fermentation. This is due to the fact that oxygen diffuses 300,000 times as rapidly in air as in water or tissue.

**Forms of fermentation.** Two forms of fermentation are illustrated in Eqs. (1) and (2), and mixed

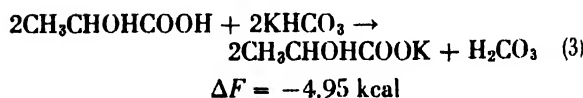
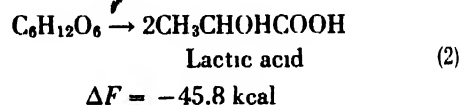
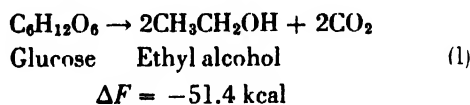
Occurrence and rate of fermentation in barley, rice, and carrot

Plant	Rate, $\mu\text{l CO}_2$ / (g wet wt)(hr)	Alcohol- $\text{CO}_2$ ratio
Barley		
Grain, soaked 12 hr	83	
Seedlings, 26 hr old	231	
Seedlings, 4 days old		
In $\text{N}_2$ 3.5 hr	521	0.33
In $\text{N}_2$ 9.0 hr	278	0.77
Rice		
Seedlings, 4 days old	224	0.98
Carrot		
Root slices	80	1.01



The influence of  $\text{O}_2$  tension on the evolution of  $\text{CO}_2$  by intact seedlings of wheat and rice. Total  $\text{CO}_2$ , respiration  $\text{CO}_2$ , and fermentation  $\text{CO}_2$  are shown (From D. L. Taylor, *Am. J. Botany*, 29(9):721–738, 1942)

reactions of alcoholic and lactic fermentation may occur. Equation (3) shows that the lactic acid may be neutralized by the buffers of the cell.



Inspection of Eq. (1) shows that alcohol is more highly reduced than sugar, whereas the carbon dioxide is more highly oxidized. Equation (2) shows that the methyl group of lactic acid is more reduced and the carboxyl group is more oxidized than the sugar. Therefore, fermentation has been known as an intramolecular oxidation-reduction. Further, Eq. (1) shows that the ratio of carbon dioxide to alcohol should be unity, and although this value is often attained, deviations from it are common. An excess of carbon dioxide over alcohol is common, as a result of nonfermentative carbon dioxide produced from the decarboxylation of organic acids (keto acids), or of the liberation of carbon dioxide from bicarbonates by acids of fermentation, or as a result of mixed fermentation with other reduced compounds in addition to, or in place of, ethyl alcohol.

**Materials used in fermentation.** Fermentation may occur at the expense of starch, glycogen, or any of several sugars. It is an enzymatic process depending upon a system of many enzymes and several coenzymes. This whole enzyme complex is

often called zymase. Nearly all the enzymes and coenzymes required for alcoholic fermentation have now been identified from extracts of higher plants (peas, beans, mung beans, and wheat embryos). Fermentation may be carried out in cell-free and concentrated plant extracts, particularly if the extracts are fortified with glucose, inorganic phosphates, and the coenzymes cocarboxylase, diphosphopyridine nucleotide, and adenosinetriphosphate. See COENZYME; ENZYME.

**Role of fermentation.** The physiological role of fermentation in higher plants is still uncertain. European and some American plant physiologists have called fermentation in higher plants anaerobic respiration or intramolecular respiration. However, these terms are disappearing from use in North America. Reactions similar to fermentation are important in the initial stages of plant respiration (see PLANT RESPIRATION). Fermentation may supply available forms of energy during germination of seeds or in plant parts infiltrated with water. With the exception of a few higher plants, for example, rice (*Oryza sativum*), all growth ceases at low oxygen pressures, but fermentation may still supply energy for maintenance of steady states and synthesis essential for survival. Anaerobic bacteria and yeasts may grow in the absence of air. See FERMENTATION; PLANT METABOLISM; YEAST [D.R.G.]

**Bibliography:** J. Bonner, *Plant Biochemistry*, 1950; W. O. James, *Plant Respiration*, 1953; F. C. Steward (ed.), *Plant Physiology*, vols. 1 and 2 1959

## Plant geography

That major subdivision of botany concerned with all aspects of the spatial distribution of plants. This science is also known as phytogeography, phytocorology, or geographical botany. By tradition, it also involves some aspects of the distribution of plants in time, or historical plant geography, paleoecology, and paleobotany. In its historical development, plant geography has been intimately connected with the rise of evolution, ecology, and genetics. It is not yet consistently segregated from ecology. From a second and equally logical viewpoint, plant geography is a major subdivision of the science of geography, although by custom few geographers deal mainly with plants. The word geobotany, undesirable etymologically, is a confusing term because of numerous and contradictory usages.

The function of plant geography is to record the observed, empirical facts of plant distribution, and also to understand and interpret these facts. Where possible, the study includes the prediction and control of distributional phenomena, especially as these relate to plant pests and to the introduction and spread of desirable species and vegetation types. Such practical aspects are pertinent to the fields of forestry, agriculture, range and pasture management, wildlife habitat management, horticulture, and soil and water conservation. Plant geography is, essentially, not an experimental science,

and in general does not involve laboratory procedures and technological equipment.

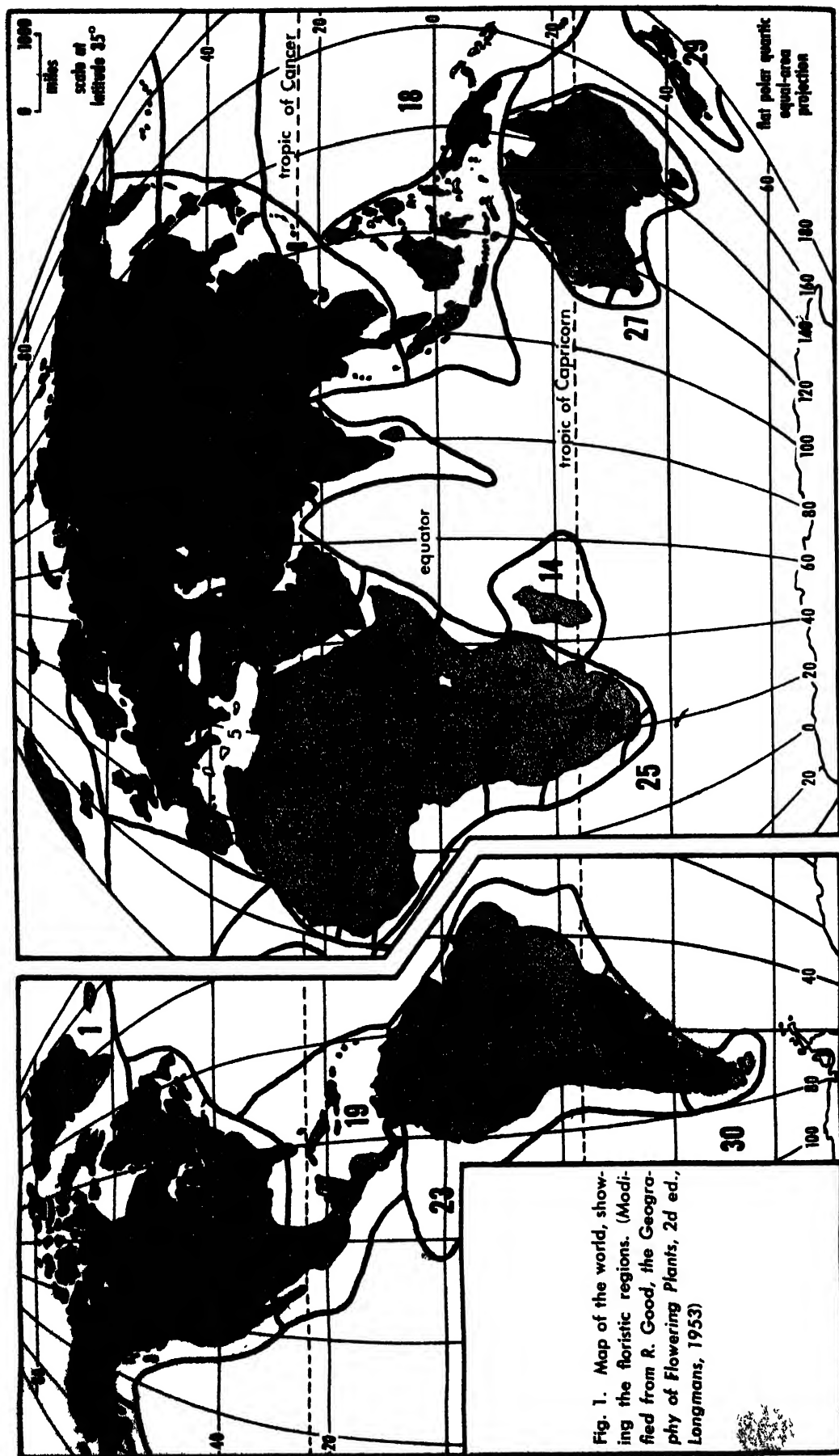
**Flora and vegetation.** There are two major subdivisions of plant geography paralleling one logical breakdown of botany itself. Floristic plant geography embraces the spatial distribution of the flora while vegetational plant geography is the spatial distribution of the vegetation. A clear understanding of the two terms is essential.

Flora is a scientific term, with no common usage. The flora of an area or period of time is the totality of all the species within that geographical unit, independent of their relative abundances and their relationships to each other. The technical term population, in this connection, refers collectively to all the individuals of any one species within a locality.

Vegetation is a term of popular origin, and refers to the mass of plant life that forms the natural or seminatural landscape. The vegetation of a region is the tapestry or carpet of plant life, developed by differential and varying combinations and growths of the numerous elements of the local flora. Technically, it is an organized and integrated whole, at a higher level of integration than the separate species, composed of those species and their populations. Sometimes vegetation is very weakly integrated, as the pioneer plants of an abandoned field. Sometimes it is highly integrated, as in the tropical rain forest. Vegetation possesses emergent properties not necessarily found in the species themselves, and is referred to by nonbiologists as a type of "organism," a different and more inclusive term than the organism of biologists.

**Floristic plant geography.** The basic components of any flora are the kinds of plants composing it, commonly referred to as species. The species can be grouped into various kinds of floral elements which are not mutually exclusive. For example, a genetic element has a common evolutionary origin; a migration element has a common route of entry into the territory; a historical element is distinct in terms of some past event; and an ecological element is related to an environmental preference. Aliens, escapes, and very wide-spread species are given special treatment. An endemic species is restricted to an area, usually small and of some special interest. See POPULATION DISPERSAL.

The idea of area is fundamental to the science and is itself the subject of a specialized section called areography. An area is the entire region of distribution or occurrence of any species, element, or even an entire flora. The local distribution within the area as a whole, as that of a swamp shrub, is the topography of that area. Areas are of interest in regard to their general size and shape, the nature of the margins, whether they are continuous or disjunct, and in their relationships to other areas. Groups of areas are unicentric or polycentric when they segregate into one or several geographically distinct territories. Areas of closely related plants that are mutually exclusive are said to be vicarious. A relict area is one surviving from an earlier and more extensive occurrence.



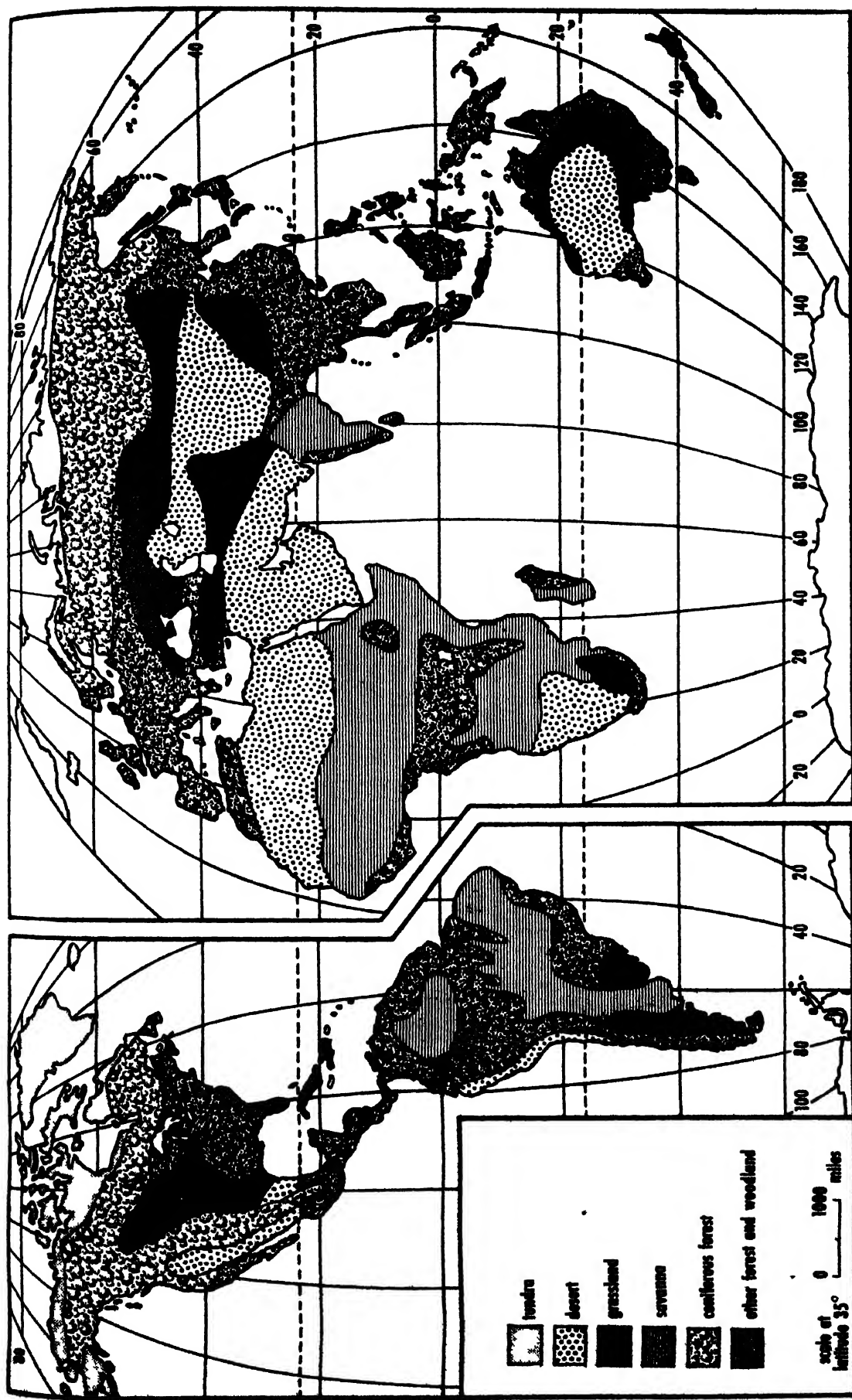


Fig. 2. Map of the world, showing the distribution of physiognomic vegetation types. (After Brockmann-Jerosch, modified from R. Good, *The Geography of Flowering Plants*, 2d ed., Longmans, 1953)

On the basis of areas, and their floristic relationships, the surface of the earth is divided into floristic regions (Fig. 1), each with a distinctive flora.

The understanding and interpretation of floras and of their distribution have been predominantly in terms of their history and ecology. Historical factors, in addition to the evolution of the species themselves, include consideration of theories of continental drift, land bridges, and orographic and climatic changes in geologic time that have affected migrations and perpetuation of floras. Ecological factors, more amenable to observation, and thus, unfortunately, to post hoc reasoning, include the contemporary roles played by precipitation, humidity, water levels, temperature, wind, soil, animals, and man. See PALEOBOTANY.

**Vegetational plant geography.** The basic components of the vegetation of any landscape are the plant communities. The science of plant communities is known as plant ecology, in the American sense, also as plant sociology, vegetation science, and phytocoenology. Many definitions of the plant community have been attempted, but none has gained universal acceptance. In part, this problem is inherent in the nature of the community itself, which is a natural phenomenon composed of elements, the species, which themselves usually maintain a high degree of independence. Thus, the community is often only a relative continuity in nature, bounded by a relative discontinuity, as judged by competent botanists. See COMMUNITY.

Vegetational plant geography has emphasized the mapping of so-called vegetation regions, and the interpretation of these in terms of environmental, or ecological, influences.

There are many aspects of a mosaic of plant communities which could serve to identify a geographic unit of vegetation, but that which has been predominant in the literature had its origins in folk knowledge. It is the physiognomic distinction between grassland, forest, and desert, with such variants as woodland (open forest), savanna (scattered trees in grassland), and scrubland (dominantly shrubs). Within forest, the chief breakdown has been into coniferous evergreen forest, broadleaved deciduous forest and broadleaved evergreen forests, mostly tropical. Furthermore, the attempt is made to map original virgin vegetation as opposed to cover types obviously due to the influence of man (Fig. 2). There is increasing dissatisfaction with this approach, but no accepted alternatives have arisen. Dissatisfaction arises from improved understanding of virgin vegetation, frequently found to be influenced by ancient human populations. Furthermore, the segregation of coniferous from deciduous types is found to separate vegetations closely related in all other aspects, such as yellow birch and hemlock in North America, and to unite types otherwise unrelated, like the pine stands which are found from the tropics to the Tundra edge. In addition, disturbance of grassland may allow the invasion of apparently self-perpetuating woody vegetation, or vice versa, in a manner that makes a physiognomic clas-

sification less fundamental. Unlike floristic botany, where evolution provides a single unifying principle for classification, the nature of vegetation in its geographical distribution is such that many types of regions and many types of classifications may have equal significance in rationalizing the natural phenomena. See VEGETATION ZONES.

The interpretation of the distribution of vegetation has been overwhelmingly in terms of the existing average environment. Catastrophic factors, such as fires, hurricanes, droughts, and other abnormal weather, are receiving increasing attention. There has been relatively little emphasis on differences due to the genetic nature of the species. For example, bristle cone pine trees have a life span of 4000 years, and Australian eucalypts were absent from, but by nature amenable to, the environmentally similar but treeless California chaparral region. From one viewpoint, it is the varying genetic demands of the different species upon their environment which permits their segregation into communities. The fact that arboreta and botanical gardens are so successful in growing many species outside their normal ranges is being recognized as a refutation of the more extreme environmentalist views.

The uniformitarian environmentalist interpretation of vegetation regions is the most completely documented. Climate is considered of primary importance. Numerous empirical formulas, combining various features of temperature and moisture, have been derived so as to correlate with the distribution of physiognomic vegetation types. Soil is recognized as secondary in importance. In addition, biotic factors, including both man and other animals, have limiting effects. Although analysis of the normal environment is essential to the full understanding of the distribution of vegetation types, it is not likely that, except for trigger factors, direct and simple cause-and-effect relationships will be found between vegetation types and those elements of the total environment which man isolates and studies. See ECOLOGY, HUMAN; ECOLOGY, PHYSIOLOGICAL; TERRESTRIAL ECOSYSTEM; see also POSTGLACIAL VEGETATION AND CLIMATE. [F.E.E.]

**Bibliography:** S. A. Cain, *Foundations of Plant Geography*, 1944; A. Engler, *Das Pflanzenreich*, 1900; A. Engler, *Die Vegetation der Erde*, 1896; R. Good, *The Geography of Flowering Plants*, 2d ed., 1953; E. V. Wulff, *An Introduction to Historical Plant Geography*, 1943.

## Plant growth

An irreversible increase in cell numbers and cell size in plants. In contrast, growth in animals is almost wholly the result of increase in cell numbers. Another important difference in growth between plants and animals is that animals are determinate in growth and reach a final size before they are mature and start to reproduce. Plants have indeterminate growth and as long as they live continue to add new organs and tissues. In animals growth of the different parts of the body is more or less simultaneous; in plants growth is restricted to the grow-

ing points or meristems (see MERISTEM, APICAL; MERISTEM, LATERAL). Therefore, in an animal most body cells have about the same age and the individual dies as a unit, but in a plant new cells are produced all the time, and some parts, such as leaves and flowers, may die, while the main body of the plant persists and continues to grow. However, the basic process of growth, cell multiplication, is very much the same in plants and animals, and mitosis and cell division will not be considered here in detail. See CELL DIVISION; MITOSIS.

This article discusses the various phenomena of growth, followed by sections on reproductive growth, dominance and germination, periodicity and abscission, and seasonal thermoperiodicity.

**Factors affecting plant growth.** The factors which control plant growth can be separated into three groups. First are the inherent genetic factors, the genes, carried by all cells, which give every cell the potentiality to grow, and which control the limits within which each cell, each organ, or each plant can develop (see GENETICS). These can be controlled only by breeding, and once the egg cell is fertilized, the genetic potentialities of the future plant are fixed. See REPRODUCTION, PLANT.

The second group of factors which control plant growth is the internal factors, such as the interactions between cells, the hormonal control system, the internal food distribution, and all correlations in general. They will be discussed later in the section on plant hormones.

The third group of factors comprises the root and aerial environment.

The balance between the available water in the soil and the water loss through transpiration is one of the major factors in plant growth (see PLANT, WATER RELATIONS OF). Another major one is the availability of nutrients (see PLANT, MINERAL NUTRITION OF). This is a function of the concentration of nutrients in the soil, and also depends on such factors as the soil water and air content, the size of the root system, the presence of a proper soil microflora, the temperature of the soil, and the balance between the nutrients. See ROOT (BOTANY).

For land plants the aerial environment is of paramount importance. The components of this environment are temperature, light, humidity, wind, air composition, and such extreme conditions as frost, excessive heat, rain, extremes of barometric pressure, invisible radiation, periodic changes, and the living plants and animals in the immediate surroundings.

It is important to realize how complex the interrelationships between external and internal environment are. The effects are both direct and indirect, for a change in growth rate or in development influences the subsequent behavior of the plant as well. At present, knowledge about the interrelationship between these factors is only fragmentary. This is partly because of the difficulty of investigating these effects. In studying this relationship it is necessary to control the external environment and to work with genetically uniform plant material.

This is now being done in so-called phytotrons, in which each of the environmental factors can be controlled separately.

The additional problem of the irregularities and unpredictability of the climate arises in dealing with the effects of climate on plant growth. The growth responses of plants to individual factors will be discussed in this article.

**Plant hormones.** The cell is the smallest unit of the organism which has all the attributes of life and which can persist as a unit. See CELL (BIOLOGICAL). It has been shown that not only the fertilized egg cell, but also isolated cells from a tissue can grow and develop into a complete plant. This was already known about certain leaf cells which, upon regeneration, give rise to plantlets on the excised leaf. Therefore, if not all, at least a considerable number of cells in the body of a plant, when isolated, can give rise to a complete plant. However, as long as these cells are in contact with their neighboring cells in the intact plant, they do not exhibit their full potentialities but remain only a part of the whole. This means that there is a controlling mechanism inside the plant which integrates the individual cells, the tissues, and the organs into a complete organism.

It has been proved in a number of cases that the interrelations between the cells which mold them into an organism are due to minute amounts of chemicals produced in one part of the plant body which activate or inhibit other parts after the chemical has been transported to them. Such chemicals, produced in small amounts in one part of the body and regulating other parts, are called hormones. In animals, most hormones are carried in the blood stream; in plants, they are transported through the living cells or in the transpiration stream.

If a cell is completely self-sufficient and can produce all the organic substances it needs for growth and metabolism, such a cell cannot be part of an organism. Many unicellular algae can grow in this way, although they may develop into irregular masses of cells (colonies) with or without specific shape. See ALGAE; CYANOPHYTA.

In yeast, a slight modification of this growth pattern exists. These cells can grow provided the concentration of some of the substances they produce themselves, such as biotin, and thiamin, is high enough in the liquid in which they grow (see YEAST). Thus, a single cell of yeast will not grow by itself, but is dependent on many other cells in its surroundings. Once enough biotin and thiamin is produced by all the cells, they will start to grow collectively.

An organism results when part of the cells have lost the ability to produce certain essential substances which are still produced by other cells of the same organism. This is perhaps best illustrated in the case of root growth.

When a root tip 1 cm long is cut off a tomato plant and placed aseptically in a medium consisting of a 2% sugar solution to which the necessary



mineral salts are added, such a root tip will grow only a small amount. But when thiamin in a concentration of 1 part in 100,000,000 is added to the medium, growth is rapid. Also, tips of the cultured root when cut off and placed in fresh medium, will likewise continue to grow. Therefore, this medium contains all that is necessary for continued root growth. In the normal plant the salts are supplied by the soil, the sugar by the photosynthesizing leaves, and the thiamin by the younger leaves. *See LEAF (BOTANY); PHOTOSYNTHESIS*. Thus the roots cannot grow faster than the production of sugar and thiamin in the leaves allows and, as a result, balanced growth occurs.

In all excised roots investigated, addition of thiamin is essential for continued growth. In other roots additional substances, such as vitamin B<sub>9</sub> and niacin, are also essential (*see VITAMIN*). Thus, thiamin satisfies the definition of a plant hormone, and vitamin B<sub>9</sub> and niacin may also be hormones in such cases as pea roots.

There are many other hormones involved in the growth of a flowering plant. For example, if the stem tip is cut off, the stem underneath stops growing in length. But when the stem tip is replaced, or when auxin, the stem-growth hormone, is applied instead of the stem tip, growth of the stem is resumed. The amounts of auxin produced by the stem tip and required for normal growth of the stem itself are infinitesimal (about  $10^{-6}$  mg) and this places auxin in the category of hormones. *See AUXIN*.

**Growth correlations.** A number of the correlations which exist in plants and which are usually influenced by hormones are shown in Fig. 1. Young leaves grow because purines, such as adenine, are supplied to them by mature leaves. Similarly an unknown growth factor, produced by roots, is essential for stem growth. It is not known where gibberellin or kinetin (a recently discovered plant hormone) fit into this scheme, but it is clear that the growth of the different plant organs is intimately interrelated through the need of some plant parts for substances produced in other parts of the plant (*see GIBBERELLIN*). Often these substances are simple compounds, and may be effective by being prosthetic groups of enzymes, and the latter could not function without such groups. *See ENZYME*.

In an unexpectedly large number of cases, the correlations in plants are produced by one simple substance, called auxin, which was first discovered as the plant-growth hormone which causes elongation of cells. Now another substance has been found, gibberellin, which more spectacularly influences elongation of stems, leaf stalks, and leaves. It is present in plant organs, but its role in normal growth has not been established.

Other correlations, effected by auxin are (1) control of lateral bud inhibition by the apical bud; (2) prevention of abscission of a leaf petiole or of a peduncle of a flower as long as auxin is produced by the leaf, flower, or fruit; (3) root for-

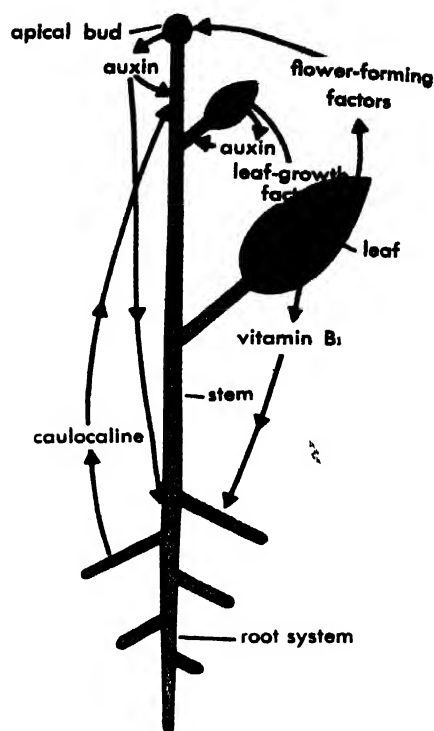


Fig. 1. Schematic drawing of a plant with root system, stem, old and young leaf, and apical bud. Some known growth correlations drawn with broken lines which indicate factors responsible for these correlations. Sugar interrelations not shown. (From F R Moulton, ed., *The Cell and Protoplasm*, Science Press, 1940)

mation at the basal end of cuttings; (4) tissue growth, and induction of cancerous growth; (5) induction of xylem elements in regenerating vascular bundles; and (6) production of parthenocarpic fruits. *See BUD (BOTANY); CANCER (BIOLOGY); FLOWER (BOTANY); FRUIT (BOTANY); VASCULAR BUNDLES*.

To explain how auxin can take part in so many reactions and can control so many correlations, it has been assumed that auxin controls one master reaction involved in each of the above-mentioned processes. This master reaction has been variously assumed to be a change in cellular permeability, a step in cellular respiration, a synthetic process, or a stimulation of translocation of other substances, but thus far no such universal master reaction has been found. Researchers are still looking for the mechanism of each of the above-mentioned processes.

**Embryonic growth.** After fertilization of the egg cell by one of the sperms produced by the generative nucleus of the pollen tube, the resulting zygote starts to develop. This occurs either immediately after fertilization, as in most rapidly developing seeds, or after a delay of several months as in the case of pines. *See SEED (BOTANY)*. During the first 5-10 cell divisions an undifferentiated mass of more or less globular tissue is produced within the endosperm. In some seeds, such as those

of orchids, development does not proceed immediately beyond this point. Even after germination of orchid seeds, the globular cell mass continues to grow evenly in all directions until a body one to several millimeters in diameter is produced, which is called the protocorm. Differentiation of stem and root is delayed until the protocorm has reached this size, which takes several months. See STFM (BOTANY).

In most plants the undifferentiated cell mass derived from the zygote proceeds directly to differentiate into an embryo consisting of a growing stem tip bearing two or more leaf primordia at one end and a root primordium at the other. Food for the growth and differentiation of the embryo is supplied by the mother plant either directly from the cotyledons, or indirectly by means of the endosperm. It has been found that the fully developed embryo can be excised and cultivated aseptically on a medium containing sugar and mineral salts. However, an immature embryo needs in addition small amounts of organic nutrients, such as vitamins. A still smaller embryo in which only the beginning of the cotyledons is indicated will not grow when excised unless it is supplied with coconut milk, a very rich source of organic nutrients, being a liquid endosperm itself.

As soon as the embryo inside the seed is fully grown it passes into a dormant condition which is broken only upon germination. The most remarkable manner in which the embryo differs from all other developmental stages of the plant is its ability to withstand almost complete dehydration, a condition which causes death at any other stage in the life of the higher plant.

**Vegetative meristematic activity.** The stem primordium of the embryo, upon germination of the seed immediately becomes the growing stem tip which continues to produce more stem cells and leaf primordia. There is a remarkable control of the cell divisions in this growing point, because the sequence of leaves follows a perfect order (phyllotaxis), which in its regularity compares with that of the structure of a crystal.

Little is known about control of the cell divisions in the growing point. In general long days are usually required to keep the cells dividing in this growing point, especially in shrubs and trees of temperate regions. It has not been possible to change the growth pattern of the stem growing point, or apical meristem, with chemicals. Microsurgery, the science of microdissection, has shown that the youngest leaf primordia have an influence on the location of the subsequent primordia. This means that the cells in the growing point, although remarkably autonomous in their growth, are to some extent dependent upon each other. This is perhaps best indicated by the existence of sectorial and periclinal chimaeras (growth distortions). When two different plant species are grafted together, a new plant occasionally develops on the junction of their tissues which combines the properties of both. This happens when a new grow-

ing point regenerates at the graft union which is comprised of cells of both species. If the two halves of the growing point are different, a plant results in which the two species are joined lengthwise. An example of such a sectorial chimaera is shown in Fig. 2. Although there is a considerable difference in size and growth rate between the two species composing this sectorial chimaera—for example, a nightshade and a tomato—when in direct contact with each other, the cells in the growing point and all along the stem grow and act in unison; that is, they divide and elongate at exactly the same rate.

Perhaps even more remarkable are the periclinal chimaeras. They have been studied in greatest detail in the case of nightshade and tomato. Occasionally a bud develops on the graft union of these two species that produces a plant in which the characters of both species are perfectly blended and in which the leaves, flowers, and fruits are intermediate between the two species in size, form, and color. No actual fusion of the nuclei of the two species has occurred; this is shown by the fact that a seed produced by these intermediate forms develops either into a nightshade or a tomato plant. Through chromosome counts and other evidence it was found that in these periclinal chimaeras the different layers of tissue, such as the epidermis, the cortex, and the central cylinder (stele), belong to the two different species (see CORTEX, PLANT; EPIDERMIS, PLANT; STELE). Either the epidermis, or epidermis and cortex, belong to one species, and the cortex and central cylinder, or central cylinder alone, belong to the other species. This is possible because in the growing point there are three layers of cells: the dermatogen which gives rise to the

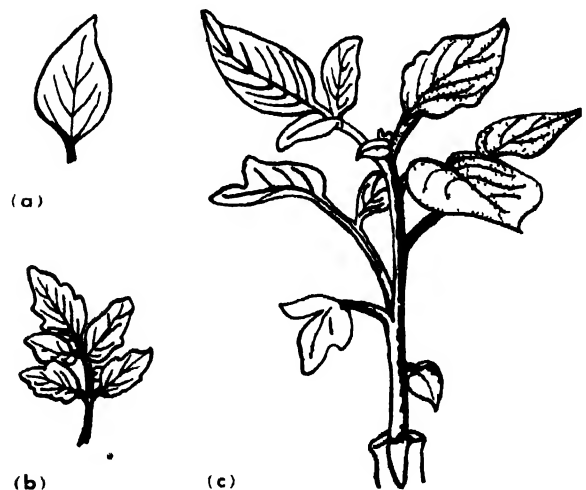


Fig. 2. A sectorial chimaera. (a) Leaf of black nightshade, *Solanum nigrum*. (b) Leaf of tomato, *Solanum lycopersicum*. (c) Base shows a tomato stem into which a wedge of nightshade tissue was grafted. After union of the two tissues, a shoot developed which consisted of nightshade tissue (right) and of tomato tissue (left). (After Winkler)

epidermis; the periblem which produces the cortical cells (outside the pericycle); and the plerome which gives rise to all the cells of the center (stele) of the stem from the pericycle inward (see PERICYCLE). All four possible periclinal chimaeras between nightshade and tomato have been obtained in extensive grafting experiments. Such periclinal chimaeras are also known for hawthorne, *Mespilus*, *Cytisus*, and *Laburnum*, and in each case intermediate forms are produced. This means that all cells and tissues influence each other during their stages of growth so that each partakes of the form and size of the other. In general, the cells produced by the periblem have the greatest effect on the final shape of leaves and flowers.

Similar periclinal chimaeras were produced in *Datura* by colchicine treatment of the growing point (see COLCHICINE). This causes formation of tetraploid cells. When only the dermatogen or periblem or plerome cells of the growing point are tetraploid, and the others are normally diploid, periclinal chimaeras are formed in which the difference in cell size does not result in distorted plants, but gives rise to harmonious structures.

Occasionally a growing point enlarges beyond its normal size. This can be stimulated by auxin treatments. It then gives rise to fasciated (flattened malformed) stems with abnormal phyllotaxis (leaf arrangement on the stem).

The stem growing point can be cultured in vitro by placing it on an agar medium containing sugar and salts. It will then develop into a stem with leaves which, upon regeneration of roots, will become a complete plant. Growing points of monocotyledons and ferns (which groups usually do not regenerate easily) are easier to culture than growing points of dicotyledons (which have a much greater capacity for regeneration).

In addition to the apical meristems there are the lateral meristems in the plant called the cambium and the cork cambium, and the adventitious growing points. About the physiology of the cambium, it is known that its activity is under the control of the buds and leaves of the plant. In deciduous trees divisions in the cambium, giving rise to phloem and xylem cells, start as soon as the buds become active in spring (see PHLOEM; XYLEM). During the period of bud opening and young-leaf expansion, the elements of the spring wood are formed. Later the presence of mature leaves causes the cambium to produce summer wood with smaller cells, thicker cell walls, and often fewer vessels. This periodicity in cambial activity gives rise to the annual xylem rings in the wood. This is not an autonomous periodicity, for when the leaves are removed the cambial activity stops. If in the same year more buds and leaves develop, new spring wood is produced, followed by summer wood, and an extra complete annual ring is formed.

It has been found that auxin application to a stem can stimulate cambial activity in the same way growing buds and young leaves do. Therefore, it is likely that the spring wood is formed under the

influence of auxin coming from the buds and young leaves. Actually it has been found that the auxin production of these growing plant parts shows the same cycle as that of cambial activity. Also in most cases cambial activity starts in the younger branches, and is followed by cambial growth in the trunk, with a time lag which approximates the rate of auxin transport from buds through branches to the stem base.

It has been possible to grow cambium cells in vitro. In most cases, however, they regenerate undifferentiated callus cells, indicating that the stimulus for cambial growth is normally more complex than merely an auxin supply. This is also indicated by the periodic changes in the cells differentiating from the cambium. For example, in a rubber tree, *Hevea*, at intervals of 1-4 months, a layer of latex vessels is produced in the phloem, alternating with layers of sieve tubes and other phloem elements. Or in the xylem of members of the plant family Sapotaceae layers of wood parenchyma alternate with layers of vessels and fibers. See PARENCHYMA.

### REPRODUCTIVE GROWTH

Reproduction in plants can either be sexual or vegetative (asexual). Some reproduction mechanisms which exist in higher plants are discussed here.

In plants such as *Ficaria*, some lilies, or *Hydrocharis*, many axillary buds swell and store food in their bud scales. These buds may then become detached, and behave like seeds; that is, when conditions become favorable, these buds start to grow and form new plants. In *Hydrocharis* or *Stratiotes* the buds are formed in autumn, as a response to the shortening of the days, and they "germinate" only after having been subjected to cold during winter (see PHOTOPERIODISM IN PLANTS). In many bulbous plants, such as tulips, hyacinths, daffodils, and onions, reproduction is mostly vegetative. The bulb consists of a markedly foreshortened stem on which thick scale leaves are implanted, and when completely developed, the stem usually ends in a flower. Toward the end of the season, two or more axillary buds on the foreshortened stem develop into new bulbs. This causes the clustering of bulbs on these plants. When the scale leaves of these bulbs are halved by cutting off the upper half, a large number of adventitious buds may form on the cut surface, each one producing a tiny bulb which can grow into another plant.

In a number of plants, thickened buds, or bulblets, are formed in place of flowers. Some plants, such as *Polygonum viviparum*, reproduce entirely in this way. In others, such as *Agave*, only occasionally do flowers fail to develop, but the flowers are replaced by such vegetative bulblets. The factors causing this transformation are unknown.

Very peculiar forms of vegetative reproduction exist in the genus *Bryophyllum*. The Madagascan species *B. tubiflorum* and *B. daigremontianum* produce, under long-day conditions, adventitious buds in the notches of the teeth of their leaves.

These buds drop off when they have formed their first two pairs of leaves, root, and form new plants. In *B. calycinum* these buds are formed in the leaf notches only after the leaf is cut off.

A large number of plants, for example, iris and lotus, form rootstocks which branch and form aboveground shoots at their apices. As the older portion of the rootstock dies, the original plant becomes divided into a number of individual rootstocks and consequently multiplies. When the rhizomes remain aboveground, they are called runners. Strawberries and many grasses multiply this way. Runner formation in the strawberry is strictly a long day response and occurs at higher temperatures.

In potatoes underground runners enlarge at the end producing tubers. This tuber formation is the result of a stimulus (or perhaps hormone) which comes from the shoot and is produced only at low temperature or in short days. This stimulus can pass through a graft union. The tubers which are formed are dormant, and to sprout, require either a chemical treatment with ethylene chlorohydrin, or several months of dormancy at room temperature. This ensures vegetative reproduction; the plants survive the cold winter as dormant tubers and new shoots develop the next spring. In many other plants, such as dahlia and sweet potato, roots swell in the same manner as the underground shoot tips of the potato and growing points on these roots develop into new shoots the following season.

Many agricultural crops are reproduced vegetatively. In addition to potatoes, sweet potatoes, and strawberries there is sugar cane which grows from stem cuttings. A cutting is a piece of stem with one or more nodes, which, when placed under suitable conditions (moist, usually partly buried in soil, sand or peat), will produce roots and shoots so that a complete plant results. Many other plants, such as Lombardy poplar and many shrubs and trees grown by horticulturists, are propagated by cuttings (see STEM CUTTINGS). The roots which develop on these cuttings either arise from pre-formed root primordia, present near the node, as in sugar cane, or are new formations, mostly near the basal cut end of the cutting. This root formation is an expression of the tendency of a part of a plant to regenerate lost organs. The hormones, necessary for root formation and root growth, are formed in the shoots and usually flow downward to the roots. When this downward transport is interrupted by cutting the stem, these hormones accumulate near the cut surface where they stimulate new formation of roots. It was found that in the majority of cases the stimulus for root formation coming from the plant top can be replaced by auxin; thus, auxin application at the base of the cutting is very commonly practiced by horticulturists to ensure good rooting. The auxin in most instances speeds up rooting on a large proportion of the cuttings. In some cases vitamins or amino acids were also found to stimulate root formation. See AMINO ACIDS.

In cases in which cuttings will not root, plants may be propagated by layering or budding.

**Floral initiation and development.** At a certain moment in the development of a vegetative growing point the regular sequence of vegetative cell divisions is interrupted and the growing point is transformed into a flower primordium. This change is induced by a stimulus, received by the leaves, and transmitted through the phloem of the stem. It can now be stated that when a purely vegetative scion is grafted on a photoperiodically induced stock, the scion will under certain conditions, such as the removal of the larger leaves from scion, start to flower. See GRAFTING OF PLANTS.

Morphologically, the first change in the transformed growing point is a broadening. Then, instead of the regular leaf phyllotaxis, the sepals are laid down simultaneously, usually in a whorl. After a time interval of a few days, another whorl of primordia, the petals, are formed. Then the stamens and carpels are produced. In exceptional cases this sequence is partly reversed; for example, in the case of inflorescences the flower primordia are produced acropetally (from the base upward) or basipetally (from the apex downward) on the large meristematic dome. See INFLORESCENCE.

The initiation of flower primordia is not exclusively under the control of the photoperiod. In many cases, as in beet and peach, flower initiation occurs only after a cold treatment called vernalization. This is an induction phenomenon, but there may be an interval of one or several months between exposure to cold and transformation of the growing point. When during this induction period the plant is subjected to high temperatures, occasionally devernization occurs with resultant non-flowering. In many plants, such as deciduous trees, coffee, and the dove-orchid (*Dendrobium crumenatum*), the flower buds develop only to a certain size and then become dormant. It then takes exposure to relatively low temperatures to break the dormancy of these buds. In peaches, pears, and other deciduous trees the buds need a period of weeks or months below 40°F before they will open. That means that they have to pass through a winter for best flowering. In the dove-orchid a rapid cooling from 80° to 65°F will make the dormant flowerbuds enlarge, and exactly 9 days later they open. In coffee the first heavy rain of the season usually triggers the flowerbud development, and 8 days later all trees which have received the rain will be festooned with flowers.

**Flower biology.** The opening and closing of flowers and the process of pollination is a fascinating field of study. Whereas many flowers, such as roses, camellias, snapdragons, and orchids, open only once and then remain open for the rest of the life of the flower, many other flowers, such as tulips, poppies, tobacco, and gazania, open and close several times with a daily rhythm. On clear days the California poppy opens its flowers at 10 A.M., and the petals close again at 4 P.M. On the same days

the tobacco, *Nicotiana affinis*, opens its flowers at sunset and closes them the next day at 9 A.M. The flower movements of the California poppy are induced by the light-dark change of the previous day; they close 21 hours after the previous sunset. The flowers of the night-blooming cactus open 24 hours after darkening on the previous day. Tulips and crocuses open upon a rise in temperature and close again upon cooling; gazania flowers open above 18°C, and close below this temperature. These responses to light and temperature are so regular that the time of day or the air temperature can be told fairly accurately by observing which flowers are open or closed. It is actually possible to use flowers as clocks. Carolus Linnaeus, the great Swedish botanist, planted part of his garden as a flower clock.

The life of flowers varies from a fraction of a day to a month or more. There are several factors which control the life span. In many flowers the petals are abscised (dropped), as in poppies and pelargoniums; in others they wilt, as in cacti and orchids. Wilting is the result of loss of vitality of the cells of the petals either because their sugars have been respired, as in *Chrysanthemum* or because their proteins have decomposed, as in cacti (see PLANT RESPIRATION; PROTEIN). In others, such as orchids, the auxin released by the pollinia causes wilting.

Most flowers have mechanisms for cross pollination; many are actually self-sterile. Sterility in flowers may be the result of a large number of factors, such as genetic block (especially in triploids and others where abnormal meiosis occurs), lack of germination of pollen, insufficient growth in length of pollen tubes, as in the long styles of *Oenothera* or corn, or lack of chemotropism, which is attraction of the pollen tubes to the ovules.

Pollination is most commonly effected by flying insects which are attracted to the flowers by color or smell. In many flowers the development of odor is periodic. In the night-blooming jasmine and in tobacco, the flowers open at sunset and at the same time become fragrant; the odor disappears again the following morning at about the time the flowers close. Other pollinators are ants, hummingbirds, or bats.

A large group of plants, such as ragweed, grasses, conifers, and trees with male flowers in catkins, is pollinated by wind; their flowers are usually inconspicuous. Their pollen, discharged in large amounts, is a major cause of hayfever.

Pollination by mechanical means is sometimes required when the pollen does not readily come out of the anthers, for example, the shaking of the flowers of tomato. Very rare are cases of pollination by water (the male flowers of *Vallisneria* float on the water and thus their open anthers come in contact with the stigmata of the female flowers). In a few cases the pollen is ejected by the snapping open of the flowers, as in *Pilea*.

Most interesting are the honey-excreting nectaries (see SECRETORY STRUCTURES, PLANT). These

are usually found at the base of the flower or in spurs (over 1 ft long in *Angraecum sesquipedale*), or sometimes outside the flower on bracts, as in the plant family Marcgraviaceae or in the genus *Euphorbia*. This honey is excreted by special glandular cells or it is pressed out of the phloem. The pollen in insect-pollinated flowers is usually sticky because it is covered with a waxlike material.

Many flowers show marked movements before opening or after pollination, for example, poppies and *Linaria*. These movements are usually due to differential growth of the two sides of the flower stalk caused by auxin. See PLANT MOVEMENTS.

**Fruit development.** Auxin plays a very important role in the growth of the ovary after pollination. The pollen itself provides the auxin, or the auxin is produced in the fertilized ovules. Therefore, an ovary of a nonpollinated flower usually does not enlarge, but if treated with auxin, it may produce parthenocarpic fruits, such as seedless tomatoes or watermelons. It was found that in the naturally parthenocarpic fruit of the navel orange, the pulp in the developing fruit provides the auxin necessary for growth. Auxin application will only lead to the formation of parthenocarpic fruit at low temperatures; during the middle of summer it has no effect.

If every flower grew into a fruit, most plants would not be able to support the crop. However many flowers drop off before fruits start to grow and some of the growing fruits also fall. Auxin sprays are sometimes effective in preventing fruit drop; at present such sprays are generally applied to prevent the preharvest drop of apples.

As a fruit grows, its total respiration increases until it is almost ripe, then respiration decreases. In some fruits, for example, the avocado, a so called climacteric rise in respiration occurs when the carbon dioxide production more than doubles just prior to the fully ripe stage; however, this happens only after the fruit is picked off the tree.

#### DOMINANCE AND GERMINATION

Every leaf carries a bud in its axil which can grow into a shoot with leaves and buds. Likewise each of these buds in its turn has the potential of growing into a leafy shoot. Should this happen, the plant would soon become an inextricable clump of branches. Although it does not occur naturally, there are certain diseases which cause every axillary bud to grow. For example, a fungus causes the development of masses of small branches called witches' brooms in trees.

**Apical dominance.** In a normal plant only a small percentage of the axillary buds grow into a shoot. When for one reason or another the apical bud ceases growth, a few axillary buds lower down the stem will develop. This suppressing effect of the apical bud on lateral or axillary buds is called apical dominance. For some time it had been suspected that the apical bud exerted its influence through the production of a hormonelike substance. After auxin was discovered to be the agent

elongation hormone produced in the stem tip, it was also shown to be the material produced by the apical bud that is responsible for inhibition of axillary bud growth. Removal of the apical bud releases this inhibition, but replacement of the apical bud with an adequate auxin supply restores the lateral bud inhibition. However, only a few hours' interruption of the auxin supply releases the lateral bud inhibition irreversibly.

Apical dominance has many interesting characteristics. For instance, the inhibition increases with distance from the tip and the buds farthest removed from the apical bud are most inhibited, whereas the buds nearest to it have the best chance for development. Upon removal of the apical bud, the uppermost axillary buds start to grow; once they start growing they inhibit buds further down from the stem tip.

In nature conditions occur under which the basal buds grow more than those nearer the apex. For example, after a warm winter peaches show delayed foliation; that is, the apical buds fail to open in spring and buds lower down the stems are the only ones which grow.

Many hypotheses have been offered to explain the mechanism of auxin action in lateral bud inhibition. The direct-action hypothesis states that this is merely a matter of auxin concentration; at low concentration auxin accelerates stem growth; at high concentration it inhibits growth. However, this does not account for the all-or-none nature of inhibition. Besides it has been shown that it is not the total amount of auxin, but the relative amounts in bud and stem that control growth. Only when the auxin concentration in the bud exceeds that in the adjacent stem is bud growth possible. In a somewhat different light, the hypotheses of indirect action assume that auxin releases inhibitors, or controls transport of other essential growth factors.

Apical dominance can appear in other ways. In conifers the perpendicular growth of the primary shoot is called orthotropic, whereas all lateral shoots grow more or less horizontally, or plagiotropically. When the apex is removed, or the apical bud is injured, one or more of the higher side-branches will become orthotropic and take over the vertical position and function of the original apical bud. This occurs in pines and firs, but in *Araucaria* the plagiotropic growth of the lateral shoots is not a response to the dominance of the apical bud, for they remain plagiotropic even in the absence of an apical bud.

**Bud dormancy.** As explained previously, the majority of buds do not grow because of the correlative inhibition established by the growth of the apical bud. As soon as this correlative inhibition is removed, some of the axillary buds begin to grow, usually within a few days. Many buds, however, do not start to grow when the growing conditions become favorable and when apical dominance is removed because such buds are dormant. Dormant buds occur on most deciduous plants in winter.

When a lilac, oak, or peach branch is cut off a plant growing outside and brought into the greenhouse, the buds will not develop between October and February. In April, or even in early summer, these plants have nondormant buds, and they start to grow as soon as the branches are removed and the cut ends put in a moist medium. Dormancy of the buds is induced by the short-day light exposure of the branches in autumn. When such deciduous trees are kept throughout the summer under long-day conditions, their buds do not become dormant and they will develop at any time when the temperature is favorable and the correlative inhibition is removed.

Induced bud dormancy can be broken in several ways. In nature it is done by exposure for several months to near-freezing temperatures. Peach or pear trees brought into the greenhouse in late summer or autumn are dormant, and they can remain dormant for more than a year because the buds have not been exposed to cold. The bud itself must be kept cold. On a branch which has been partially cooled, only the buds on the cooled part develop.

In a number of cases a chemical treatment can substitute for the cold requirement, and ether or ethylene chlorohydrin vapors as well as nitrophenol and ethylene have been used to break bud dormancy. Chemical treatments have the advantage of requiring only a few days in contrast to the cold exposure which must last weeks or months.

Bulbs and tubers commonly have to go through a dormant period before they can sprout. In some cases this is a question of straight dormancy (gladiolus, potato, lily-of-the-valley) in which cold or ethylene chlorohydrin will break it. Other cases will be discussed in this article in the section on seasonal thermoperiodicity.

**Germination.** In the section on embryonic growth the formation of the seed was described, and it was stated that upon ripening, the seed passes into a dormant condition, from which it emerges upon germination.

The seed represents a stage in the development of a plant which is particularly resistant to cold, heat, and drought. The main function of a seed is to provide for progeny by carrying the plant over unfavorable conditions and facilitating distribution of the species.

In germination three distinct stages can be distinguished. The first stage comprises the intake of water, a stage that is completed when all cell walls and protoplasts have sufficient water content. Associated with the water uptake is an increase in respiration. The second stage is a curious one. Except for respiration, no measurable changes occur, the embryo does not enlarge, and the seed seems to be in a condition of suspended animation. During these first two stages germination is a reversible process; the seeds can be dried and rewetted a number of times without any effect on their future germinability. Large numbers of seeds persist for years or even decades in the soil without reaching the third stage of actual enlargement of the em-



bryo. However, once this third stage has set in, there is no holding back and the embryo goes through its exponential growth to form the seedling, or it dies.

The second stage, that of suspended animation, is the most critical because it is the period when it is decided whether the seed will germinate or not; it is an all-or-none process. No matter how long or short this second stage of germination lasts, once the inhibition has been released, growth of the seedling will always be the same.

Some of the mechanisms involved in preventing growth of the embryo and the events occurring during the second stage of germination can now be discussed.

For a number of seeds it has been established that they contain substances which inhibit germination. When such inhibitors are removed, either by decomposition or by leaching, embryo growth proceeds normally. In at least one seed, *Amaranthus*, and probably in many others, such inhibitors also repress respiration. Thus, *Amaranthus* seeds can be kept for many years in moist soil without germinating or utilizing all their storage food.

In other seeds it has been found that the seed coats, or certain layers of them, are impermeable either to water or to oxygen. Breaking (scarifying) of such impermeable layers leads to germination.

In many leguminous seeds in which the seed coat is so hard that the embryo cannot break through, scarification suffices for germination.

In the seeds of wild plants, a large number of mechanisms delay germination until the growing conditions for the seedling are favorable. Cultivated plants are not sown until the season is proper for germination and growth; consequently, germination delays are unnecessary and may be unfavorable. The germination behavior of seeds of cultivated and wild plants should be clearly distinguished.

Many seeds fail to develop and ripen under conditions which seem unusually favorable for germination, such as the moist interior of a melon or of a tomato fruit. In such species there are substances in the fruit pulp which inhibit germination. In exceptional cases, however, germinated seeds are found inside the fruit, for example, occasionally in oranges or peaches.

Some seeds of northern crop plants, such as barley and wheat, germinate at low temperatures and can be sown in autumn or early spring. Others, such as corn, sugar beet, and tobacco, require higher temperatures and are sown much later in spring, or they are sown in greenhouses and the young seedlings are transplanted in the field, for example, tomato and chili pepper. The latter plants originated in warm climates and grow only during the summer.

Seeds of tropical forest trees usually have a short life, and unless they germinate immediately, they decompose in the humus layer.

Many seeds which normally remain dormant during winter need stratification, that is, treatment with low temperatures (near freezing), before they will germinate. This treatment is effective only when the seeds are moist.

Most of the seeds which have inhibitors, hard or impermeable seedcoats, or need stratification before they can germinate, have normal embryos which will start to grow as soon as conditions are favorable. There are other seeds the embryos of which will not develop even when excised. Such embryos are either immature or have embryo dormancy. In both cases germination can occur after a sufficient period of waiting during which after ripening (physiological change) takes place.

The seeds of many plants germinate very irregularly, only a few at a time. In this way germination is sometimes spread over a period of many years, with occasional peaks of germination at about yearly intervals.

Seed germination of desert plants has been studied in some detail. These seeds show a number of different mechanisms which prevent premature germination. When rains are few and far between the chance that a small amount of rain will be augmented by another rain is slight. Under these conditions germination in desert plants would best be delayed until about 1 in. of rain has fallen continuously, and when the soil is sufficiently wetted to ensure normal development of the seedling. Some seeds have four or more mechanisms for delayed germination. Among these belong (1) presence of germination inhibitors which can be removed by leaching, as occurs during a heavy rain; (2) excess mineral salts in the seedcoat or in soil, also removed by heavy rain; (3) remains of fruit covering, which has to be removed before seed can germinate; and (4) hard seedcoat, which may be removed by scarification, usually by rubbing of the seeds with sand or stones; for example, after a heavy rain a slurry of water, sand, and gravel rushes down the wash carrying with it the seeds of smoke tree, palo verde, and other shrubs.

Growing plants have a strong inhibitory effect on seeds present in the soil around them. Germination of many plants does not occur in a closed vegetation. Interesting exceptions are seeds of parasitic or semiparasitic plants, such as *Striga* and *Euphrasia*. Their seeds germinate only under the influence of excretions from the roots of their host plants. The reverse phenomenon was found for the guayule rubber plant, *Parthenium argentatum*, whose seedlings are killed by excretions, consisting partly of cinnamic acid, from the roots of the mature plants. See CARBOXYLIC ACID.

#### PERIODICITY AND ABSCISSION

There are relatively few plants whose growth is not periodic. Basically all processes in the growing point are periodic; for example, the laying down of leaf primordia is cyclic and occurs in a definite sequence producing a specific phyllotaxy.

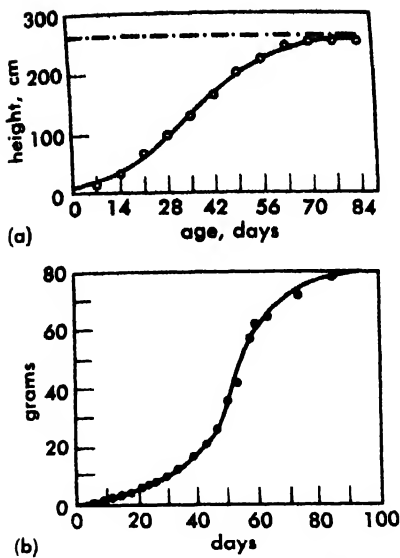


Fig. 3. (a) Growth in height of sunflower stem (from E. P. Odum, *Fundamentals of Ecology*, Saunders, 1953). (b) Growth of entire corn plant as measured by increase in dry weight (from Gustav Backman after Stefanowska, in D. W. Thompson, *On Growth and Form*, Cambridge, 1942).

**Growth periodicity.** Each individual cell and each internode go through a succession of growth rates, the so-called grand period of growth. When the length of a cell or of an internode is recorded as a function of time, a sigmoid (*s*) curve is obtained, showing that after an initial slow start, growth increases to a maximal rate, and then gradually decreases (Fig. 3). This is typical for the individual cell, for a colony of cells, such as a bacterial or yeast culture, and even for a whole organism. Because a growing plant usually has the same number of dividing, growing, and maturing cells in its stem, the growth rate of a whole stem is regular and linear, always maintaining the same proportion of growing cells in their different stages. Thus, the regular growth rate of the whole plant is the integration of thousands of sigmoid growth curves of the individual cells.

As was explained in the section on dormancy, the periodicity of growth of a pear or a peach tree is partially induced by the external conditions. But there is an inherent periodicity in growth of a number of plants, often based on the morphology of the plant. In tulips and other bulbous plants the growing point goes through a cycle of forming scale leaves, true leaves, and a flower. Because the flower terminates the growing point, a new axillary bud takes over, and a new cycle is started. In the tomato plant the same phenomenon occurs. After forming 8–17 nodes (depending on the variety), the growing point transforms into a terminal flower. Then a lateral bud starts to grow, forms 2–4 nodes, and likewise terminates in a flower. Thus the tomato stem is a sympodium (many forked), and a definite periodicity in development can be ob-

served. In pines, peaches, and most other trees there is a sequence of formation of nodes bearing leaves and bud scales. Even if there is no cessation of growth when the bud with bud scales is formed, for example, because the bud did not become dormant as a response to short-day treatment, the cycle of scale and leaf formation is repeated several times per year, giving evidence of an internal growth periodicity.

Many cases of periodicity in development have been shown to be induced by external factors. The daily fluctuation of growth rates of stems, roots, and leaves has been found to be due, at least in part, to the daily light-dark cycle, to the periodicity in temperature, or to the change in relative humidity from day to night. However, in the nyctinastic ("sleep") movements of leaves, there is also an autonomous component in these growth periodicities which has a 24-hour cycle and becomes synchronized with the 24-hour climatic cycle.

Very curious periodicities are known in the flowering of orchids, coffee trees, and other plants. For example, the dove-orchid, *Dendrobium crumenatum*, flowers in the lowlands of Java once every few months, but all plants in the same locality flower on the same day. This was found to be the result of a sudden drop in temperature accompanying certain heavy rains.

The simultaneous flowering of all bamboo plants of a particular species every 11 or 33 years has tentatively been connected with the 11-year cycle in sunspots. But other plants, such as *Strobilanthes* in Ceylon and Java, flower simultaneously at 4- or 7-year intervals. Such periodicities seem to be of the same type as those of the 13- or 17-year cicada.

Typical yearly periodicities in growth and flowering are often induced by the yearly cycle of temperatures or photoperiods.

**Abscission in plants.** Because special tissues are produced between stem and leaf stalk, or between stem and fruit stalk, or at the base of petals, each plant does not remain burdened with its dying

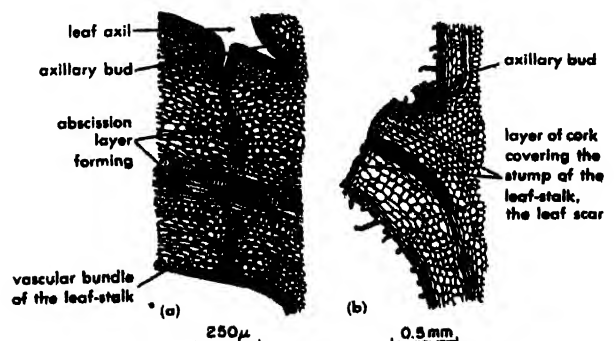


Fig. 4. Abscission. (a) Longitudinal section through base of a leaf of *Prunus*. Divisions of cells form an abscission layer, shown vertically. (b) Vertical section through part of a stem and leafbase of *Coleus*, after abscission of the leaf. (From R. D. Gibbs, *Botany, An Evolutionary Approach*, Blakiston-McGraw-Hill, 1950)

and dead leaves, flowers, or fruits (Fig. 4). These dead parts drop off (abscise) and are regenerated to carbon dioxide in the carbon cycle of nature often before the whole plant has died and fallen to the ground.

Although not much is known about how or why a leaf abscises, a great deal has been learned about how to stimulate or prevent abscission. The abscission layer at the base of the petiole is formed when the leaf blade is removed, or when the leaf is no longer active, as in autumn. The formation of this abscission layer can be delayed by applying auxin, or other substances with similar physiological activity, on the leaf stalk stump.

The reaction of the ovary and fruit is similar to the leaf. When the ovules in the ovary have not been fertilized, no auxin reaches the place of attachment of the fruit stalk to the stem, and the young fruit abscises. The lack of the inhibitory effect associated with developing seeds can be replaced by auxin application to the ovary. This means that auxin production by the developing seeds has at least a dual function: causing the growth of the ovary, and preventing its abscission as explained in the preceding section on fruit development.

There are several practical applications of these auxin effects. One is the spraying of ripening apples with auxins to prevent preharvest drop, and also to prevent bruising of the fruit as they fall on the ground. Another is the use of auxin sprays in tomato growing to prevent the flowers on the first clusters from dropping off, and to cause the continued growth of the young fruit.

In a number of flowers, fertilization causes abscission of all the flower parts except the ovary; this presumably is also connected with the auxin production by the developing ovules or by the auxin released by the pollen. The latter is the case in the postfloration phenomena in orchids. Instead of placing the pollinia on the stigma, auxin can be applied and will produce swelling of the ovary and wilting of the petals.

The mechanical cotton pickers work properly only when the cotton plants are leafless. Therefore, methods have been developed to defoliate cotton fields. This can be done either by killing the leaf blades, which will cause abscission of the petiole as though it had been debladed, or by applying an antiauxin.

Not only petioles, but also whole branches may be dropped through the formation of an abscission layer. This occurs in *Castilloa* and in *Sterculia*, and results in a self-pruning operation in which the older branches, shaded by the higher ones, are abscised.

**Tissue culture in plants.** Under the heading of tissue cultures, often the problem of organ culture is also considered. The growth of roots in vitro was discussed under the general heading of plant hormones, and the culture of growing points was treated under vegetative meristematic activity. Typical plant tissue cultures were achieved for the first time about 1938. When pieces of stem, root, or

other organs are placed aseptically on an agar medium containing sugar and mineral nutrients, they either regenerate buds and roots and grow as a complete plant, or they produce a small mass of undifferentiated tissue called callus, which develops into a globular mass on top of the piece of stem. To make this globular mass of undifferentiated parenchyma cells grow to a larger mass, it is necessary to add auxin to the medium.

Thus callus tissue, transplanted aseptically in a medium containing sugar, mineral salts, and auxin, will grow to a large mass of undifferentiated cells. This mass of callus can be subdivided into smaller pieces, and each piece will continue to grow equally well on this medium. Whereas this is the rule, occasionally a piece of such a callus culture will lose the requirement for auxin in the medium, and will then continue to grow on a medium containing only sugar and salts. This is called habituation, and makes these tissues less dependent upon their environment.

Usually plant tissues are completely dependent for growth upon the surrounding tissues and organs, and such tissues fail to grow when the adjacent parts become mature. An exception exists in the case of crown gall, an abnormal growth which usually occurs near the root crown on stems. Crown gall is induced by *Bacterium tumefaciens* and it is largely a mass of callus. It is possible to keep the crown gall callus growing even when the bacteria have been eliminated because the callus tissue loses the requirement for auxin in the medium. It will grow in the same sugar-nutrient medium in which habituated cultures develop. Because of its independence of auxin, it has escaped the growth control of the plant. This explains why a crown gall develops.

More recently a number of substances have been isolated and identified which increase growth of callus cells, usually manifold. Many are present in coconut milk. Among them kinetin, or 6-furfuryl-aminopurine, has been investigated most extensively.

The most interesting new development in tissue culture is that it has been found possible, by several different means, to make a single callus cell develop into a large callus mass, and to have this differentiate into a shoot and ultimately into a complete plant. Thus, a single undifferentiated cell has the potentialities of developing into the complete organism, just as the egg cell has.

Usually, no differentiation of tissues and organs occurs in a callus mass. Upon transplanting a piece of differentiated tissue into it, however, differentiation of callus cells in the immediate vicinity of the transplant will occur.

#### THERMOPERIODICITY

Temperature influences most physiological processes. The optimal temperature is the temperature at which a process is fastest or most efficient. Above the optimal temperature, the rate of the process decreases, often rapidly, as a result of injury to the protoplasm. There is also a minimum

temperature below which the process does not go on at all. In many cases this lowest temperature is  $0^{\circ}\text{C}$  or even slightly below freezing. Because of the osmotic value of the cell contents, cells do not freeze unless cooled to well below  $0^{\circ}\text{C}$  (see OSMOSIS). Above the minimum temperature, the rate of the process increases rapidly, in an exponential manner, much as temperature influences the rate of chemical processes. This means that for every rise in temperature of  $10^{\circ}\text{C}$ , the rate of the process is doubled or trebled. The rate of the process at a given temperature divided by the rate at a temperature  $10^{\circ}\text{C}$  lower is called the temperature coefficient or  $Q_{10}$ .

**Seasonal thermoperiodicity.** Many plants cannot develop normally in a constant temperature or when the daily temperature range is kept within the same limits. For example, sugar-beet plants kept every day at  $23^{\circ}\text{C}$  and every night at  $17^{\circ}\text{C}$  will continue to grow vegetatively for 3–4 years, after which time the base rots away. A peach tree can be kept for many years under the above-mentioned temperature regime without ever flowering. Tulip and hyacinth bulbs, planted in the equatorial climate of the tropics, even at higher altitudes, will not develop normally, and will die in 1–2 years. All these plants will develop normally when they are subjected to a yearly cycle of low temperatures followed by high temperatures. This requirement is called seasonal thermoperiodicity.

In tulip and other bulbs, the seasonal thermoperiodicity has been investigated in great detail. It was found that each developmental stage of the plant has its own optimal temperature. When the above-ground part of the tulip plant has died, the bulb is completely filled with storage food in the scale leaves, and the growing point has formed only a few leaf primordia. During the next few weeks, more leaf primordia and a flower primordium are produced. This process of morphogenesis has a relatively high temperature optimum of about  $25^{\circ}\text{C}$  (see PLANT MORPHOGENESIS). When the bulbs are subsequently kept at that temperature very little else will happen. Although ultimately a few weak leaves might appear, the flower-stalk never elongates. To make the plant develop normally the temperature must be lowered after the flower primordia have been initiated. By keeping the bulbs at  $0^{\circ}\text{C}$ , they can be stored for 6 months or longer without damage, following which they can be forced to flower, producing tulip flowers at any desired time.

At a somewhat higher temperature,  $5\text{--}10^{\circ}\text{C}$ , the bulbs still seem to remain dormant, but when they are afterwards brought to  $13\text{--}17^{\circ}\text{C}$ , they will sprout rapidly. This does not happen in the bulbs kept at  $0^{\circ}\text{C}$ . It is evident that pretreatment at  $5\text{--}10^{\circ}\text{C}$  for 1–2 months is essential for later growth. Thus  $5\text{--}10^{\circ}\text{C}$  is a delayed optimum in which the effect becomes apparent some time after the treatment.

After the  $5\text{--}10^{\circ}\text{C}$  treatment, the initial optimal temperature for stem elongation is about  $13^{\circ}\text{C}$  until flower opening when it shifts to  $17^{\circ}\text{C}$ . In this way the following sequence of optimal tempera-

tures has been found for tulip development, each lasting 1–2 months:  $25^{\circ}$ ,  $5\text{--}10^{\circ}$ ,  $13^{\circ}$ , and  $17^{\circ}\text{C}$ . A period of blooming and photosynthesis follows after which the same temperature sequence must be repeated.

For a hyacinth the optimal temperature sequence is  $34^{\circ}$ ,  $25^{\circ}$ ,  $13^{\circ}$ ,  $21^{\circ}\text{C}$ . In this plant each temperature is also connected with a specific phase of development. Therefore, these bulbs must pass through a sequence of different temperatures to allow the various morphogenetic and physiological processes to take place. In a tropical climate with constant temperature, the bulbs do not pass through such a sequence and hence the plants do not develop normally. In a temperate climate with the proper yearly cycle of temperatures, the bulbs not only develop properly, but they also become synchronized with the climate.

There is a large number of plants in temperate climates which have to pass through a cycle of high-low-high temperatures within a period of about 1 year before they develop normally. In beets, carrots, and other biennial plants, a low-temperature period is necessary before flower initiation can occur later at higher temperatures. In deciduous trees the lower temperatures are necessary for later flower initiation, and for breaking of bud dormancy.

**Daily thermoperiodicity.** Practically everywhere in the world the temperature during the day exceeds that during night. It has been found that plants grow better when subjected to such a daily temperature fluctuation than when exposed to a constant temperature. From the limited information available, a  $6^{\circ}\text{C}$  temperature differential between day and night seems to be optimal. In desert plants this optimal differential seems to be higher; in some tropical plants it may be lower.

Only part of the explanation of this phenomenon lies in the different processes which predominate in the plant in darkness and in light, each with its own optimal temperature. In the tomato plant, for instance, most growth occurs during the night at an optimal temperature of  $17\text{--}18^{\circ}\text{C}$ . During the day the optimal temperature is about  $23^{\circ}\text{C}$  which coincides with the higher optimum of photosynthesis.

This explanation, however, accounts for only part of the daily thermoperiodicity. In most plants it seems that the autonomous 24-hour cycle is of paramount importance for normal development. If the internal cycle does not coincide with the period of the external cycle, or if there is no external cycle to synchronize the cyclic processes inside the plant, development slows down or becomes abnormal. Thus the daily temperature cycle and the light-dark cycle tend to steer development into normal channels. See PLANT PHYSIOLOGY.

[F. W. WENT]

**Bibliography:** W. Crocker, *Growth of Plants*, 1948; W. Crocker and L. V. Barton, *Physiology of Seeds*, 1953; W. E. Loomis, *Growth and Differentiation in Plants*, 1953; F. W. Went, *Experimental Control of Plant Growth*, 1957; P. R. White, *Cultivation of Animal and Plant Cells*, 1954.

## Plant hormones

Defined by F. W. Went and K. V. Thimann (1937) to be "a substance which, being produced in any one part of the plant, is transferred to another part and there influences a specific physiological process." Hormones can be described as chemical messengers. They transmit the controls of specific processes from one part of the organism to another. They are usually present in relatively small quantities, in contrast to substrates or most nutrients. Plant hormones which control growth have been identified, and hormones controlling flowering are suspected but not yet identified.

With the agglomeration of large numbers of cells into organisms, there must be some organization of the cell mass, or else the result will be no more than a callus, colony, or tumor. In green plants, this organization involves in part a development for the promotion of photosynthesis and the elaboration of basic food materials. Such organization must entail a control of the sites of growth, of the differentiation of cells into tissues, and of the development of direction or polarity in the whole mass. Thus in a complex plant there must be regions of growth; there must be tissues which serve the various needs of water supply, food transport, photosynthetic activities, mechanical support, and the other functions of plant organisms; and finally there must be a direction of the whole organism into a top and bottom, or shoot and root. The plant growth hormones are the principal chemical messengers that are known to carry on these physiological controls. They are the developmental forces which direct, so to speak, the growth and form of the plant.

**Nomenclature.** According to the van Overbeek Committee (1954), the most generally accepted terms covering the plant growth hormones and related natural and synthetic compounds include *plant regulators*, the natural and synthetic organic compounds which in small amounts promote, inhibit, or otherwise modify any physiological process in plants; and *plant hormones*, plant regulators which occur naturally in plants and usually move within the plant from a site of production to a site of action.

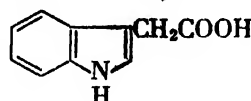
Two major divisions of plant regulators are growth regulators (compounds which affect growth) and flowering regulators (compounds which affect flowering). The same divisions apply to plant hormones; thus there are plant growth hormones, of which several are known and many are yet to be identified, and possibly flowering hormones, none of which are definitely identified as yet.

The word auxin designates one kind of growth regulator, referring to that kind which has the capacity for inducing elongation in shoot cells in the manner of indoleacetic acid, the best-known plant growth hormone. Included in this category are synthetic plant growth regulators and the plant growth hormones. The best-known synthetic auxin is 2,4-D (2,4-dichlorophenoxyacetic acid), the weed killer.

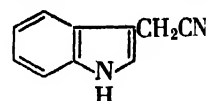
**Known plant hormones.** From the above classification, it will be discerned that two types of hormones are assumed to exist in plants, growth hormones and flowering hormones. In addition, three other types of natural growth regulators are recognized: the gibberellins, which have enormous effect on cell elongation, the kinins, which have effects on cell division, and finally growth inhibitors. Each of these three types of factors has been detected in plants, often in surprisingly large amounts for such powerful biological agents, and the identities of some of them have begun to emerge from the research of plant biochemists and physiologists. Whether or not these materials are plant hormones in that they move from one part to another is still to be learned.

Each of the two types of plant hormones will first be discussed, followed by a brief description of the three other types of natural growth regulators.

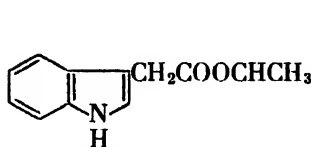
**Occurrence of growth hormone.** The commonest plant growth hormone, indoleacetic acid, has been found in many higher plants. It was first isolated from plants by K. V. Thimann (1935) who found it in bread mold, and it has since been identified in corn, oat, bean, sugar cane, pineapple, tomato, and many other plants. Until the 1950s it was thought that this compound was the growth hormone in all higher plants. However, increasing evidence is accumulating which makes clear that many other compounds are present in plants which can do the same job in stimulating growth. Only a few of these have been identified, including



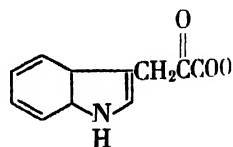
Indoleacetic acid



Indoleacetonitrile



Ethyl indoleacetate

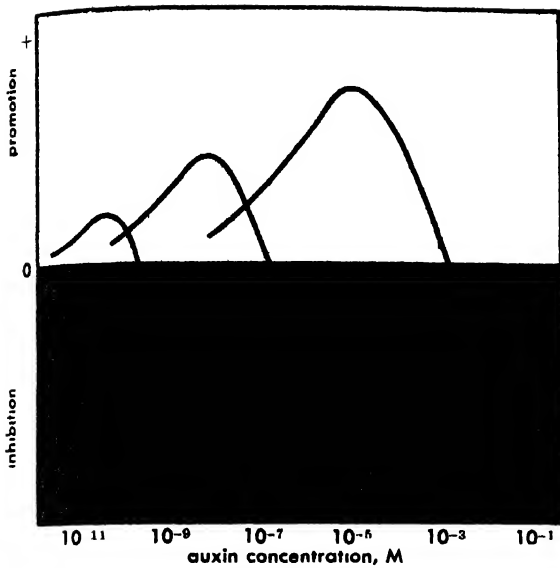


Indolepyruvic acid

indoleacetonitrile, ethyl indoleacetate, and indolepyruvic acid; formulas for these plant growth hormones are reproduced here. On chemical grounds one can see that each of these might be easily converted into indoleacetic acid, and for each of these compounds some evidence has been accrued that the conversion occurs before the stimulation of growth is achieved. However, the question is not completely answered, and numerous other compounds are yet to be identified, some of which give no chemical tests for the indole ring and hence should be something quite different.

Indoleacetic acid is formed from tryptophan. It has been suggested many times that the formation may occur via either tryptamine or indolepyruvic acid, and then indoleacetaldehyde, an immediate precursor to indoleacetic acid.

It is interesting to note that the enzyme which changes the indoleacetaldehyde to the



(growth responses of various plant organs to the plant growth hormone, indoleacetic acid. (Adapted from K. V. Thumann, 1937))

indoleacetic acid is extraordinarily sensitive to radiation damage. Radiation of plants with either neutrons or ionizing radiation, such as x-rays, causes blockage of the hormone-synthesizing activity in the plant, as pointed out by S. A. Gordon in 1956.

Once the growth hormone is formed, it moves about through the plant with considerable speed (about 1 cm/hour), and this movement has the remarkable quality of being polar; that is, the hormone moves from apex to base. The main sites of synthesis of the hormone are the tips of stems, the young expanding leaves, and other rapidly growing organs such as flowers, fruits, and root tips. The hormone formed in these sites then moves toward the base of the plant in a polar manner, and in doing so it apparently imposes many polar effects in plant development. For example, the polar movement of hormone from stem tips contributes to the inhibition of lateral buds and branches, holding them in a nongrowing condition near the top of the plant and exerting lessening inhibitions as the distance from the tip increases. As another example, the hormone formed at the tip of stems influences the angle and the rate of growth of lateral branches, thus the more auxin there is streaming from the stem tip, the more nearly horizontal will be the angle of the lower branches and the less rapid the growth of the lower branches. As yet another example, the hormone causes the differentiation of roots when it accumulates in rather large amounts, so that when a stem is wounded or cut off, interrupting the polar movement of auxin, roots will form at the site of interruption.

There must be a way for the plant to dispose of the growth hormone, and in fact there are two pathways known at present by which the hormone can be destroyed or converted to another product. Destruction of the hormone can occur by oxidation. This can come about through the action of light as an oxidant, or through the action of

ionizing radiation, or by the actions of oxidative enzymes. The best-known enzyme which oxidizes indoleacetic acid is peroxidase. Many researchers believe that the bulk of the hormone in plants is disposed of through the action of this enzyme. Plants growing in the dark are abnormally elongated in part because the absence of light results in less auxin oxidation than in normally lighted plants. The second pathway of growth hormone removal from the plant may be through the formation of addition products. These include the glycoside of indoleacetic acid and peptides, especially indoleacetyl aspartate.

**Action of growth hormone.** When the growth hormone was first discovered, it was thought that it promoted the growth of stems and leaves and inhibited the growth of roots. This was deduced from experiments in which these various plant parts were soaked in auxin solutions. However, it was later found that auxin may either promote or inhibit growth of any given plant organ, some organs being more sensitive than others. This differential sensitivity can be described in the manner shown in the illustration. Roots, buds, and leaves are all shown to increase in growth with increasing auxin concentration, but the optimum for each organ is different. Thus an auxin concentration which promotes the growth of stems can inhibit the growth of buds or roots. By this concept, the same hormone system may stimulate some growth functions in the plant and inhibit others at the same time and even at the same place in the plant.

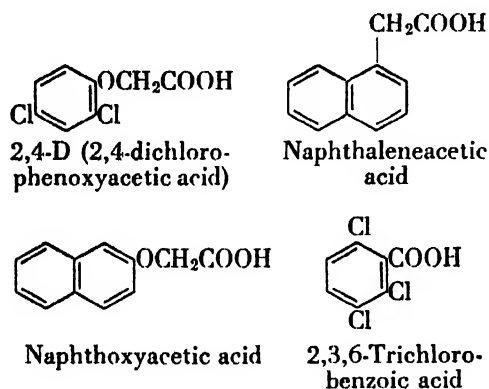
In addition to the control over growth exerted by the growth hormone there are two major effects of the hormone on plants. The hormone plays a basic role in the differentiation of tissues and organs. For example, it appears to play a major role in the differentiation of xylem, the principal water-conducting system of plants. The rates of xylem development have been closely correlated with the amounts of auxin moving through stems and young developing leaves. The stimulation of root differentiation is the best-known effect of the growth hormone on organ development.

Among the most striking effects of the growth hormone are the correlation effects. These are regulatory actions exerted by one plant part on another. An example is the suppression of growth of lateral buds by the terminal growing point of a stem, or the regulation of the branch angle or lateral growth rate. These effects collectively are called apical dominance, and the auxin produced by the growing point and associated tissues plays an intimate part in this correlation effect. Another is the regulation of abscission of leaves, flowers, or fruits; in this case, the auxin produced by the leaf blade or by the flower and fruit deters the development of a separation layer. Still another correlation effect is the formation of roots at the basal end of cuttings; in this case, the auxin produced in the stem or its buds stimulates the differentiation of roots at the basal end of cuttings where the polar translocation of the auxin has been interrupted. The correlation effects of the plant growth



hormone are classic examples of chemical means by which separate plant parts can regulate and be regulated in their growth and differentiation activities; such regulatory actions between cells are an essential ingredient for the development of multicellular organisms.

The mechanism of action of the growth hormone is an exciting area which is now receiving careful attention from plant biochemists. Several theories have been proposed to explain the action, but each has been made doubtful by further examination and experiment. Tests of various synthetic chemicals for auxin activity have permitted some generalizations about what the molecule must be like to be active in stimulating growth as an auxin. Almost all the compounds having this activity contain an aromatic nucleus and an acid side chain in a particular spatial configuration. Some synthetic compounds which are widely used as auxins include 2,4-D and related chlorinated acids, naphthaleneacetic acid, naphthoxyacetic acid, and 2,3,6-trichlorobenzoic acid. Formulas for some common synthetic auxins are given here. The manner in



which these various auxins can control the growth of plants is not clear, but the bulk of the evidence has at least established that the site of action is at the cell wall. In some manner the auxins cause a loosening and elasticizing of the cell wall, and through a stretching process, cell growth then takes place with associated deposition of new cell wall materials. The action is a metabolic one, and results in marked stimulation of the metabolic rate as growth is increased.

**Uses.** Agriculturally the auxins have found numerous uses and have led to the beginnings of chemical control of crop growth. The largest use of the auxins has been in the area of herbicides. These chemicals cause death of plants at extremely low rates of application, and also are highly selective. In general, they are more lethal against the broad-leaved plants than against the grasses, with the result that they are most useful in killing broad-leaved weeds in fields of small grains and corn. The most common auxin herbicides are 2,4-D, 2,4,5-T (2,4,5-trichlorophenoxyacetic acid), and the trichlorinated benzoic acids. In 1952, 60,000,000 lb of 2,4-D was produced in the United States alone. See HERBICIDE.

The actions of auxins on tissue differentiation

permit three other agricultural uses: the rooting of cuttings, the control of fruit abscission or drop, and the induction of flowering. Rooting can be greatly stimulated in plant cuttings by dipping the basal ends into either a solution or a dry powder of auxin (see STEM CUTTINGS). Indolebutyric and naphthaleneacetic acids are the commonest auxins for this use. The premature drop of apples and some other fruits can be prevented by the application of auxins. Weak solutions of 2,4,5-trichlorophenoxypropionic acid are sprayed onto the trees several weeks before the fruit might start to drop, and the drop is not only alleviated but the color of the apple fruits is considerably increased. The initiation of flowering can be triggered in the pineapple with the application of naphthaleneacetic acid, a fact which has been highly useful in the commercial production of these fruits, simplifying the harvest schedules enormously. Few other species of plants have been found to share this flowering response to auxins.

Among other growth functions which the plant growth hormone may regulate are the processes of fertilization and the commencement of fruit growth, or fruit set. When pollen germinates on a flower ovary, a surge of growth hormone production results, which starts the growth of the fruit from the ovary. In some species of plants, the pollen can be substituted for by the application of synthetic auxins. This chemical setting of fruit, which results in a seedless fruit, has found commercial application in the culture of tomatoes and figs.

**Flowering hormone.** The seasonal flowering of many plants has been found to be controlled by the length of day in the varying seasons. From this discovery of photoperiodism, it has been found that the stimulus to flower in response to changes in day length originates principally in the leaves of plants. Thus the stimulus must move from the leaves to the growing point where flower buds will be differentiated. This stimulus, then, can be considered as a flowering hormone, although it has not been separated from intact plants. See PHOTOPERIODISM IN PLANTS.

Various research workers have grafted together plants which have different day-length requirements for flowering, and found that when a flowering plant is grafted to a nonflowering plant in various combinations, the flowering stimulus can move over to the nonflowering member of the pair and make it flower. It is deduced that the same flowering hormone is involved in the plants of the various photoperiodic classes. The stimulus can apparently cross between plants only when there is a living connection between them, and there has been essentially no success in extraction or diffusion of a flowering hormone out of plants. The flowering hormone is not the same as the growth hormone, because it can move in any direction in living tissues, and the growth hormone moves in a polar basal direction. See GRAFTING OF PLANTS.

**Other growth regulators.** One of the most dramatic discoveries in plant chemistry was that of the gibberellins. These compounds were

Japanese scientists from rice plants infected with a disease which causes abnormally extensive growth. The fungus which causes the growth response, *Gibberella fujikuroi*, was found to produce a chemical responsible for the great growth which has been purified and identified and named gibberellin (see GIBBERELLIN). There are several closely related chemicals which occur naturally in plants and fungi, and these have profound influences in the normal growth and development of plants. They are specifically involved in the bolting or shooting up of flower stalks of plants such as spinach, tobacco, and cabbage, and they may be used commercially in causing the bolting and flowering of such species. It is beginning to be realized that the gibberellins are intimately involved in the flowering stimulus of these species, though they are not specifically the flowering hormone. The gibberellins are also effective in fruit set, and are thought to work in conjunction with the growth hormone in this physiological step. They are best known in the popular sense for the dramatic effects they have when sprayed onto plants, or they cause very large increases in plant height. For example, application to a cabbage plant can cause it to grow to a height of more than 10 ft. The gibberellins are also able to break the dormancy of many buds and seeds.

Another kind of growth regulator which is involved in the plant hormone systems is the kinin, cell division factor. Several synthetic chemicals have been found which stimulate cell division in plants. Many of these are purines related to adenine. Several materials showing this type of activity have been isolated from tissues in which abundant cell division is taking place, such as in the germinating coconut or enlarging young fruits. These chemicals act in conjunction with the growth hormone, resulting in growth by cell division instead of the usual cell enlargement response to the growth hormone. No commercial applications of the kinins have yet been developed.

Plant growth inhibitors are receiving increasing attention as growth regulators in plants, and have strong interactions with the plant growth hormone, although it is not clear yet whether they may be hormones themselves. Many growth inhibitors have been found in plants, including various lactones such as coumarin and scopoletin, various phenols such as chlorogenic acid and naringenin, and various aromatic acids such as salicylic acid (the essential part of aspirin). The plant growth inhibitors are responsible for many types of dormancy in plants, and their actions are closely interwoven with the growth hormone and the gibberellins. See AGRICULTURAL SCIENCE (PLANT); AUXIN; PLANT GROWTH; PLANT METABOLISM; PLANT PHYSIOLOGY.

[A. C. LEOPOLD]

**Bibliography:** A. C. Leopold, *Auxins and Plant Growth*, 1955; A. C. Leopold, *Plant Growth and Development*, 1964; G. Pincus and K. V. Thimann (eds.), *The Hormones*, vol. 4, 1964; B. B. Stowe and T. Yamaki, The history and physiological action of the gibberellins, *Ann. Rev. Plant Physiol.*,

8:181-216, 1957; H. B. Tukey (ed.), *Plant Regulators in Agriculture*, 1954; F. Went and K. V. Thimann, *Phytohormones*, 1937.

## Plant keys

Botanists have used analytical keys in the identifying of plants for many years. The basic principle of such a key is the finding of contrasting characters and the use of these to subdivide the group being studied into two or more branches. For example, in a collection of plants, one has compound leaves in contrast to the others having simple leaves. On this basis, the plants are separated, and the one with compound leaves is eliminated from the range of possibilities inherent in the other plants; thus the problem of identification is narrowed. The key may state that some plants are hairy and some glabrous, that some have yellow flowers and others white. By careful examination of the plant, by contrasting enough characters, and by eliminating enough members of the group, the number of possibilities is finally reduced to one, the name of the plant being identified. If the plant being studied differs in one or more characteristics from all other known plants, it is possible that a new plant has been found.

Several different types of keys have been devised. A much used type is the indented key, in which the descriptions of each character are increasingly indented a given distance from the left-hand margin of the page. The lines become more and more indented as the key is extended to include a larger

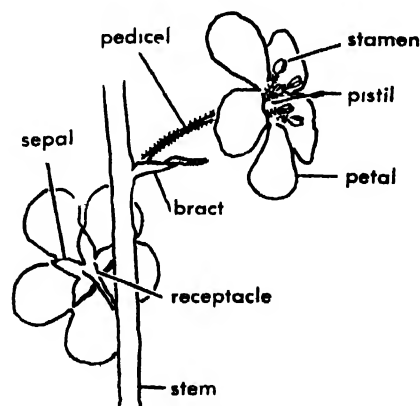


Fig. 1. Diagram of a complete flower.

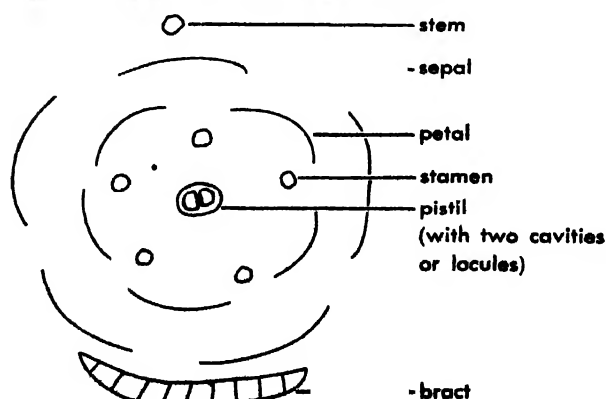


Fig. 2. Floral diagram of the flower shown in Fig. 1.

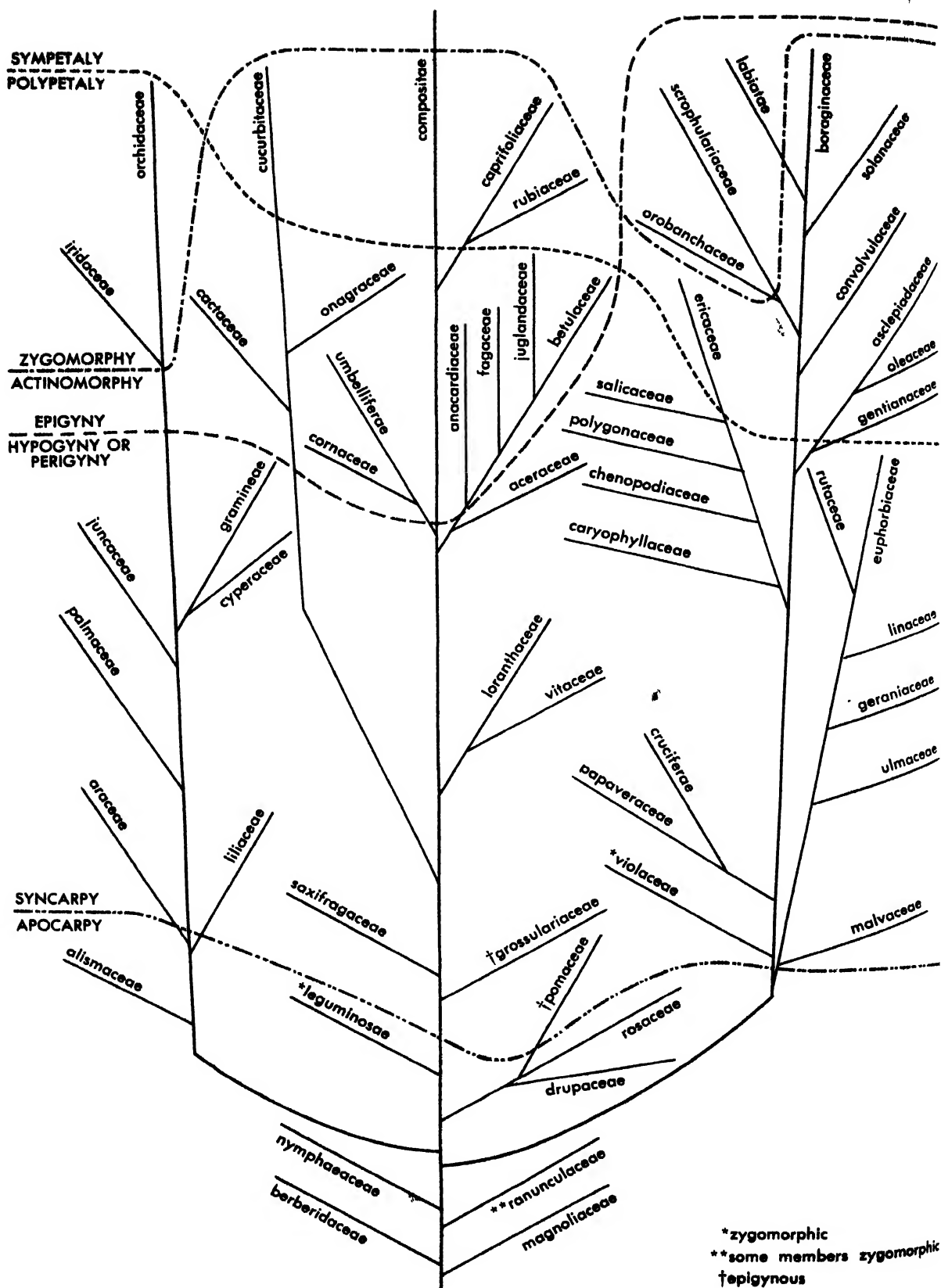


Fig. 3. Floral chart of most of the large and important families of angiosperms in the United States, ar-

ranged essentially according to the system of Charles E. Bessey.

number of categories. For an example of an indented key see Table 1. An example of a parallel key is given in Table 2.

In the parallel key two or more contrasting characters appear in consecutive lines on the page so that they can be easily compared.

When concerned with relatively small groups, in general, indented keys are preferable. When concerned with larger groups, parallel keys are more satisfactory.

Table 1. Example of an indented key (genus *Cardamine*)

a Leaves pinnately divided	
b Leaves nearly all basal	<i>C. hirsuta</i>
b Stem leafy to the inflorescence	
c Lateral leaflets oblong; plants of wet grounds	
d Leaflets of lower leaves rounded and stalked, petals 1-1.5 cm long	<i>C. pratensis</i>
d Leaflets with bases decurrent along the rachis, petals 1.5-4 mm long	<i>C. pennsylvanica</i>
c Lateral leaflets linear, plants usually of dry grounds	<i>C. arenicola</i>
Leaves entire or toothed, not pinnately divided	
b Flowers purple	<i>C. douglassii</i>
b Flowers white	
c Stem erect from a tuberous base	<i>C. bulbosa</i>
c Stem decumbent, stoloniferous	<i>C. rotundifolia</i>

Table 2. Example of a portion of a parallel key (classes and families)

Mostly herbaceous plants with vascular bundles scattered through the pith Leaves with closed venation, mostly parallel-veined, rarely net-veined Parts of flowers usually in 3s or 6s, never in 5s Herbs except in the genus <i>Smilax</i> (2)	
Herbaceous and woody plants with the vascular bundles usually arranged in a ring about the pith Leaves net-veined, with open venation Flower parts usually in 4s or 5s (8)	

## 2. MONOCOTYLEDONS

2 Flowers with or without perianth, on a thick, fleshy spadix Leaves petioled and net-veined	Araceae
2 Flowers with a perianth, not borne on a spadix (3)	
Perianth parts dry and scalelike or chaffy	Juncaceae
Perianth parts herbaceous or colored; neither scalelike nor chaffy (4)	
4 Perianth free from the ovary	Liliaceae
4 Perianth more or less adherent to the ovary (5)	
Flowers irregular; stamens 1-2, gynandrous	Orchidaceae
Flowers regular except in Iris; stamens 3-6 (6)	
6 Flowers imperfect, plants dioecious, climbing; leaves net-veined, ovate	Dioscoreaceae
6 Flowers perfect (7)	
Stamens 6, perianth parts imbricated in the bud	Amaryllidaceae
Stamens 3, perianth parts convolute in the bud	Iridaceae

Diagrams, symbols, and formulas may be included in keys. These, in a sense, are botanical shorthand, in which many features of a plant may be represented in detail in limited space.

**Floral diagram.** The floral diagram is essentially a graphic diagram of a cross section of a flower (Fig. 1), and may be represented as shown in Fig. 2. It will be noted that there are four (sometimes more, or less) sets of floral parts, arranged in whorls, one within another.

**Floral symbols.** These may be used to represent various features of a flower. Thus CA might represent the calyx (composed of sepals), CO the corolla (composed of petals), S the stamens, and P the pistil (composed of carpels). Small figures, written as exponents, may be used to indicate the number of parts, as CA<sup>5</sup>. Many other features can be indicated by additional symbols.

**Floral formulas.** Combinations of floral symbols make up floral formulas. Thus the structure of a flower might be represented by the following formula:

$$CA^3 CO^3 S^6 \underline{P}^{\circ}$$

This formula signifies that the calyx is composed of 3 sepals, the corolla of 3 petals; that there are 6 stamens, and that the pistil is made up of 3 carpels, united, as indicated by the circle about the 3. The line beneath the P indicates that the pistil is superior.

**Floral charts.** A floral chart may serve both as a key to families or orders and as a graphic representation of relationships. When concerned with relatively small groups (having few families), floral charts, like keys, are easier to make and to use. The floral chart presented in Fig. 3, arranged essentially according to the system of Charles E. Bessey, is not complete for the entire United States, but it includes most of the large and important families and illustrates the general nature of all similar charts. The left branch represents the monocotyledons, and the center and right branches the dicotyledons. The irregular, broken cross lines show that certain similar evolutionary tendencies have occurred in different groups of higher plants. For example, all families below a certain line have regular corollas (actinomorphy), whereas those above this line have irregular corollas (zygomorphy). All families below another line have carpels distinct (apocarpous), whereas all above it have them united (syncarpous). By noting that these lines cross the several branches of the phylogenetic system, it is seen that the same evolutionary changes (irregularity of flowers, union of carpels, unisexuality) took place repeatedly in different phylogenetic lines and at different times, as distinct and independent processes. Notes in the lower right corner record important exceptions, but occasional species showing deviations from the rules occur in many other families.

See PLANT CLASSIFICATION; PLANT KINGDOM; PLANT TAXONOMY. [E.L.C.]

## Plant kingdom

The world-wide array of plant life. It includes plants that have roots in the soil, that live on or within other plants and animals, that float on or swim about in water, and that are carried on dust particles in the air. There is great variation in body form and size, ranging from microscopic one-celled organisms such as bacteria and certain algae to the giant redwood trees of California, some of which exceed 300 ft in height and 30 ft in diameter and are over 3200 years of age. Growth habits and life histories also differ greatly. More than 350,000 living species of plants have been described. New discoveries add hundreds of new species each year.

This large number of forms necessitates some method of systematic plant classification so that plant scientists and others may handle effectively this great variety of organisms. A fairly complete

classification of the white oak, for example, would read as follows:

Phylum Tracheophyta  
 Subphylum Pteropsida  
 Class Angiospermae  
 Subclass Dicotyledoneae  
 Order Fagales  
 Family Fagaceae  
 Genus *Quercus*  
 Species *alba*

Scientific name, *Quercus alba*; common name, white oak

A modern system of classification of the entire plant kingdom is shown in outline form in the accompanying table.

Articles on all the listed plant groups appear under their specific names. See PLANT CLASSIFICATION; PALEOBOTANY. [P.D.S.]

### Plant kingdom

#### Subkingdom: Thallophyta

##### Division: Algae

- Phylum: Cyanophyta
- Phylum: Euglenophyta
- Phylum: Chlorophyta
- Phylum: Chrysophyta
- Phylum: Pyrrophyta
- Phylum: Phaeophyta
- Phylum: Rhodophyta

##### Division: Fungi

- Phylum: Schizomycophyta
- Phylum: Myxomycophyta
- Phylum: Eumycophyta

#### Subkingdom: Embryophyta

##### Phylum: Bryophyta (Atracheata)

- Class: Musci
- Class: Hepaticae
- Class: Anthocerotae

##### Phylum: Tracheophyta

###### Subphylum: Psilopsida

- Class: Psilophytineae
- Order: Psilophytales†
- Order: Psilotales

###### Subphylum: Lycopsida

- Class: Lycopodiaceae
- Order: Lycopodiales
- Order: Selaginellales
- Order: Lepidodendrales†
- Order: Pleuromeiales†
- Order: Isoetales

###### Subphylum: Sphenopsida

- Class: Equisetaceae
- Order: Hyeniales†
- Order: Sphenophyllales†
- Order: Equisetales

###### Subphylum: Pteropsida

- Class: Filicinae
- Order: Coenopteridales
- Order: Ophioglossales
- Order: Marattiales
- Order: Filicales

##### Class: Gymnospermae

- Subclass: Cycadophytaceae
- Order: Cycadofilicales†
- Order: Bennettitales†
- Order: Cycadales

##### Subclass: Coniferophytaceae

- Order: Cordaitales†
- Order: Ginkgoales
- Order: Coniferales
- Order: Gnetales

#### Subphylum: Pteropsida (Cont.)

##### Class: Angiospermae

##### Subclass: Monocotyledoneae

- Order: Pandanales
- Order: Helobiales
- Order: Triuridales
- Order: Graminales
- Order: Palmales
- Order: Synanthales
- Order: Arales
- Order: Farinales
- Order: Liliales
- Order: Scitaminales
- Order: Orchidales

##### Subclass: Dicotyledoneae

##### Group: Archichlamydeae

- Order: Casuarinales
- Order: Piperales
- Order: Salicales
- Order: Myricales
- Order: Balanopsidales
- Order: Leitneriales
- Order: Juglandales
- Order: Fagales
- Order: Urticales
- Order: Proteales
- Order: Santalales
- Order: Aristolochiales
- Order: Polygonales
- Order: Centrospermales
- Order: Ranales
- Order: Papaverales
- Order: Sarraceniales
- Order: Rosales
- Order: Geraniales
- Order: Sapindales
- Order: Rhamnales
- Order: Malvales
- Order: Parietales
- Order: Opuntiales
- Order: Myrtales
- Order: Umbellales

##### Group: Metachlamydeae

- Order: Ericales
- Order: Primulales
- Order: Ebenales
- Order: Gentianales
- Order: Tubiflorales
- Order: Plantaginales
- Order: Rubiales
- Order: Campanulales

† Known only as fossils; no living representatives.

## Plant metabolism

The entire complex of interrelated chemical reactions characteristic of living plants constitutes plant metabolism. Knowledge of plant metabolism has progressed through the following stages: (1) the chemical identification of the major constituents of plants and the early recognition of the processes of respiration and photosynthesis; (2) the discovery of enzymes and of the prominent enzyme-catalyzed reactions and reaction sequences; and (3) the problem of finding the function of the observed compounds, enzymes, and reactions and the means by which they are controlled to provide that orderly ongoing series of events characteristic of living organisms. None of these stages of discovery is complete. New plant constituents are reported each year. For each new compound, there exists the problem of determining its mode of synthesis and the function it serves. Each enzymatic activity found in extracts of cells must be related to the metabolism of the whole organism. The elaboration of one means of metabolic control emphasizes by contrast the inadequate understanding of others. This fragmentary knowledge nonetheless does permit the construction of an orderly concept of metabolism.

It is convenient to consider that all the reactions in living cells fall into three categories: (1) those which provide cell-stuff, (2) those which provide energy and (3) those which have no known function. In green plants the process of photosynthesis supplies from carbon dioxide and water the simple compounds from which fats, carbohydrates, proteins, nucleic acid, and other cell substances are synthesized (see PHOTOSYNTHESIS). Nongreen plants require an outside source of organic carbon. In both cases the metabolic pattern for the synthesis of cell-stuff is the same. By many series of enzymatically catalyzed reactions, a fraction of the simpler organic compounds is converted into cell substance at the expense of energy produced by the oxidation of the remainder. Because energy is required for the synthesis of each of the major constituents of living cells, an important phase of the study of metabolism is a consideration of the processes and reactions by which this energy is made available. Because there is no recognized function for several classes of substances, such as alkaloids, terpenes, tannins, and anthocyanins, which occur in plants, it is impossible to ascribe any metabolic function to the reactions which produce these compounds. See BIOLOGICAL OXIDATION.

It has been estimated that there may be several thousand reactions occurring in a living cell. Each of these reactions is catalyzed by a specific enzyme, a protein molecule with highly specific catalytic properties. Until 1958 only about 650 enzymes had been studied. Since there are no doubt many more, it is obvious that it is not possible to construct a complete picture of metabolism. An orderly concept of metabolism is possible as a result of the discoveries (1) that the complex energy-

yielding degradations of metabolites and the syntheses of new substances proceed by interdependent series of rather simple chemical changes, and (2) that the many enzymes catalyzing the metabolic reactions are not scattered at random throughout the cell but are oriented in the structure of the cell in a way that facilitates their serial action on metabolite molecules (see ENZYME; PROTEIN). The metabolic functions associated with the structures of a green plant cell are listed below.

<i>Cell part</i>	<i>Metabolic functions</i>
Chloroplasts	Photosynthesis: photolysis of water; photosynthetic phosphorylation; fixation of carbon dioxide; synthesis of carbohydrates
Mitochondria	Oxidation of various metabolites; transfer of electrons to oxygen; oxidative phosphorylation; synthesis of proteins
Nucleus	Synthesis of nucleic acids; synthesis of proteins
Endoplasmic reticulum } Ribonucleoprotein particles }	Synthesis of proteins; synthesis of nucleic acids

Many, perhaps most, of the reactions occurring in plant cells also occur in animal cells and bacterial cells. There is a striking unity in the processes by which various cells obtain and utilize energy and in the processes by which the major cell substances are synthesized. Indeed, much of the knowledge of plant metabolism has resulted from studies which simply confirmed the existence of metabolic pathways already known for animals or microorganisms. Even a superficial consideration of different organisms suggests that there must also be a uniqueness in the metabolism of each organism. See CELL (BIOLOGICAL).

## ENERGY METABOLISM

Whatever its other characteristics, life certainly requires energy. The main business of a cell is to obtain this energy and to use it in self-maintenance and in self-duplication. A living cell uses energy in performing (1) chemical syntheses, (2) osmotic work, (3) mechanical work, (4) electrical work, (5) light production, and (6) heat production. In the cells of higher plants most of this energy is produced by the aerobic process called oxidative phosphorylation. This process consists of a multi-phase enzymatic transfer of electrons from reduced cofactors through a series of cytochromes to oxygen and the concomitant conservation of chemical energy in a metastable phosphate compound. Reduced diphosphopyridine nucleotide (DPN) and flavin adenine dinucleotide (FAD) (Fig. 1) serve as donors of electrons to the cytochrome



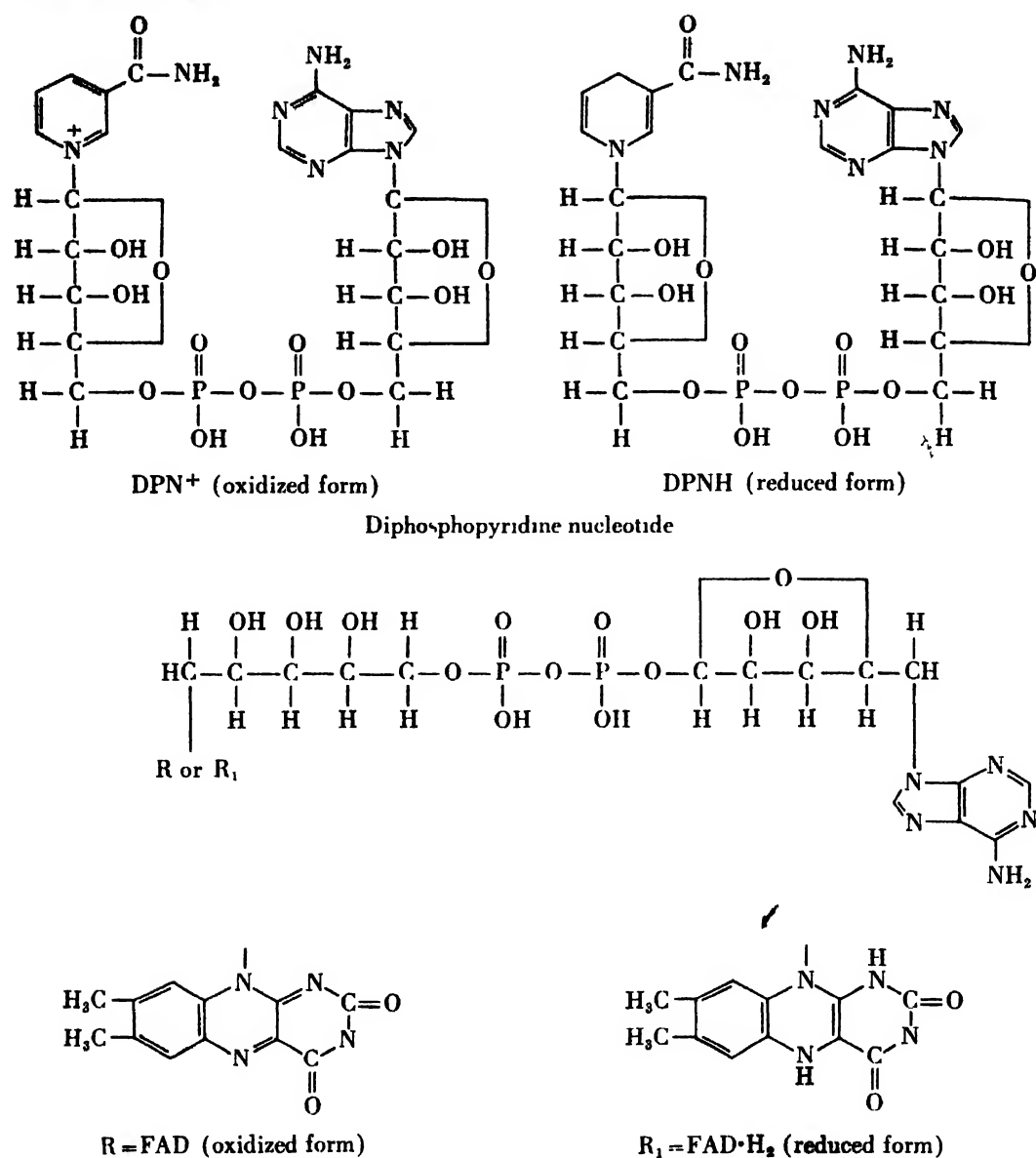


Fig. 1. Structural formulas of diphosphopyridine nucleotide and flavin adenine dinucleotide.

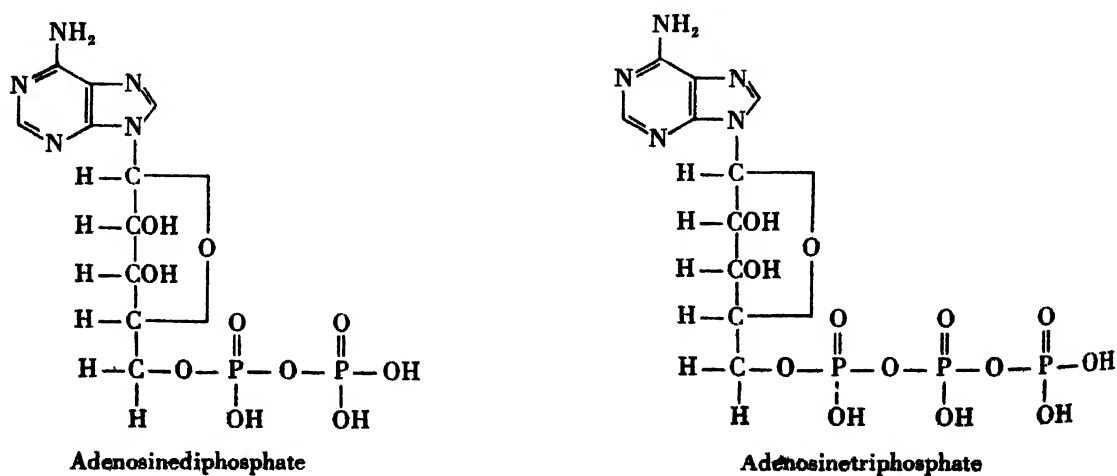
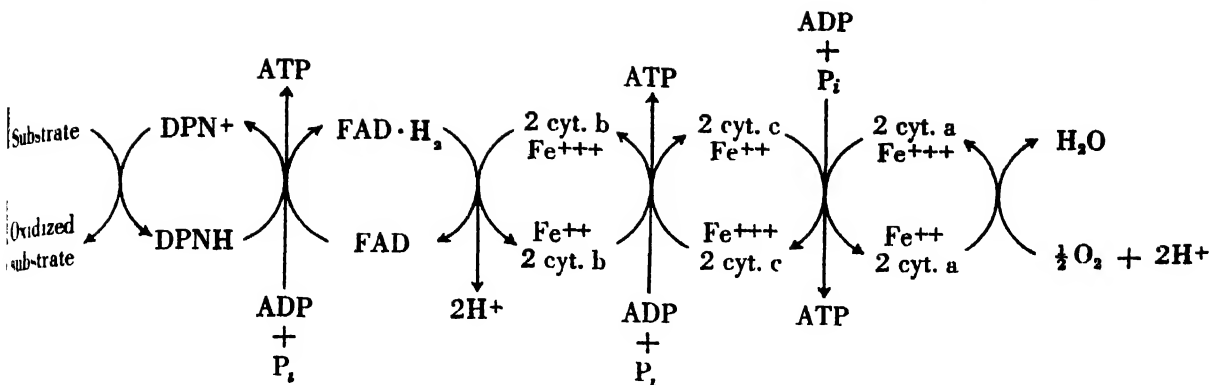


Fig. 2. Structural formulas of adenosinediphosphate and adenosinetriphosphate.



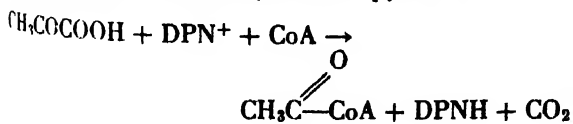
3 Electron transport and oxidative phosphorylation.

The cell components involved in this electron transfer system as well as those required for the simultaneous phosphorylation are all localized in the mitochondria of the cell. The phosphorylation which occurs during electron transport to oxygen is the uptake of inorganic phosphate and the formation of one additional pyrophosphate bond (Fig. 2) on adenosinediphosphate (ADP) to form adenosinetriphosphate (ATP). The scheme of known reactions in electron transport and oxidative phosphorylation is shown in Fig. 3.

The discovery of this important means of synthesis of ATP is of significance because ATP is the immediate source of the chemical energy required to perform the work functions listed above. Cleavage of either of the pyrophosphate bonds in ATP results in the release of about 8000 cal/mole. The enzymatic apparatus of living cells allows them to leave these bonds in a way that makes this energy available for energy-requiring syntheses and for other kinds of work.

The production of reduced pyridine and flavin nucleotides also occurs primarily in the mitochondria by a series of dehydrogenation reactions in which these enzyme cofactors are reduced during the oxidation of pyruvic acid to carbon dioxide and acetic acid, and the further oxidation of acetic acid to carbon dioxide and water through a cyclic series of reactions, the citric acid cycle (Fig. 4).

The first step in this series of oxidative reactions is the oxidative decarboxylation of pyruvate.



Acetyl coenzyme A (CoA) then condenses with oxalacetate to form citrate and free coenzyme A, which can be used again in the formation of another molecule of acetyl coenzyme A from pyruvate. The reduced cofactor (DPNH) is reoxidized by transfer of electrons through the cytochrome system. From the standpoint of energy metabolism, the importance of the citric acid cycle is based upon the production of reduced pyridine nucleotides which, on reoxidation, supply the energy for

formation of adenosinetriphosphate. A small fraction of the energy required by plant cells is provided by anaerobic processes as discussed under carbohydrates.

The metabolism of different cell substances is interrelated and interdependent in many ways. For example, the cofactors required by the citric acid cycle enzymes are also required in other oxidation-reduction reactions, such as the synthesis and degradation of fatty acids and of carbohydrates. Therefore, there is an interdependence of the reactions catalyzed by these enzymes. In addition a particular compound may be an intermediate in the metabolism of several different substances. Acetate, as acetyl coenzyme A, is produced from pyruvate and from the oxidation of fatty acids and is, in turn, a precursor of fatty acids, certain organic acids, and isoprenoid compounds. Furthermore, because adenosinetriphosphate is the immediate source of energy for most energy-requiring processes, the relative rates of these processes will be a function of the level of adenosinetriphosphate. The rate of production of adenosinetriphosphate by oxidative phosphorylation is partly a function of the concentration levels of adenosinediphosphate and inorganic phosphate end products in the utilization of adenosinetriphosphate. These interrelationships of the different reactions in metabolism, although incompletely understood, undoubtedly constitute a system of automatic controls that permit cellular metabolism to proceed smoothly. See ADENOSINEDIPHOSPHATE (ADP); ADENOSINETRIPHOSPHATE (ATP); METABOLISM.

#### CARBOHYDRATE METABOLISM

The synthesis of carbohydrates from carbon dioxide and water by the process of photosynthesis is one of the main features of carbohydrate metabolism of green plants. These carbohydrates then provide the carbon skeletons for all the other cell substances. Although phosphorylated derivatives of glyceraldehyde, erythrose, xylulose, ribose, ribulose, fructose, and sedoheptulose are involved in the cyclic process of carbon dioxide fixation in photosynthesis, much of the carbon fixed is transformed into fructose, glucose, sucrose, and starch.

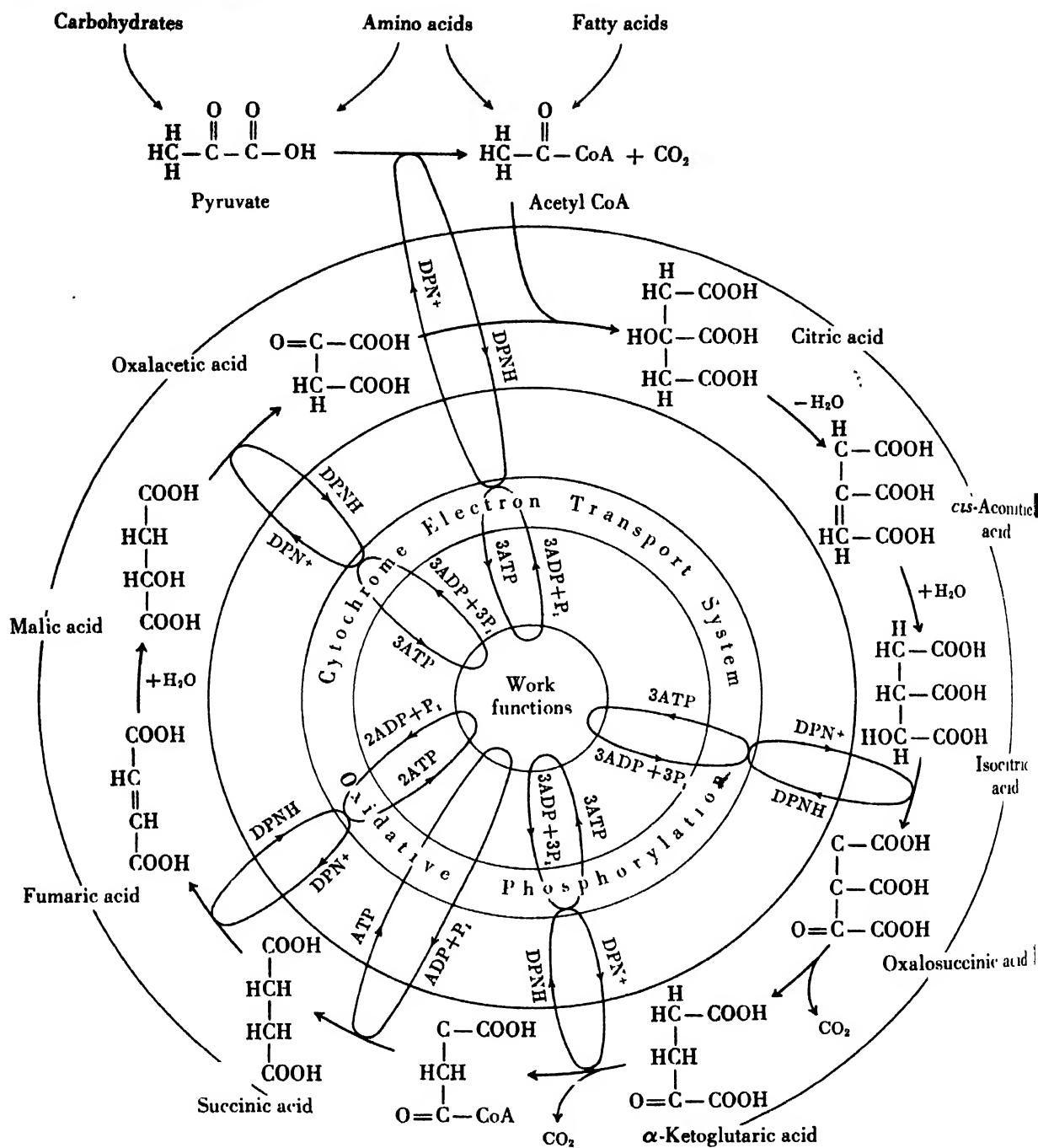


Fig. 4. Citric acid cycle.

These carbohydrates then undergo further metabolism independent of the process of photosynthesis.

The glyceraldehyde-3-phosphate produced in the photosynthetic cycle gives rise to other carbohydrates, as shown in the reactions in Fig. 5.

When glucose (from photosynthesis or from hydrolysis of the storage carbohydrates, starch and sucrose) and fructose (from sucrose or inulin) are degraded to obtain energy or to provide carbon skeletons for the synthesis of other compounds, the initial reactions involve a series of phosphorylated intermediates. This pathway of dissimilation (Fig.

6) is referred to as glycolysis or the Embden-Meyerhof-Parnas pathway. Under anaerobic conditions appreciable quantities of ethanol accumulate and there is a net production of only 2 moles of ATP for each mole of glucose converted to ethanol and carbon dioxide. Although many microorganisms can obtain sufficient energy by anaerobic processes to maintain their life cycles, higher plants cannot. Under aerobic conditions the DPNH produced by the oxidation of phosphoglyceraldehyde to phosphoglyceric acid is reoxidized by the cytochrome system and ethanol is not formed.

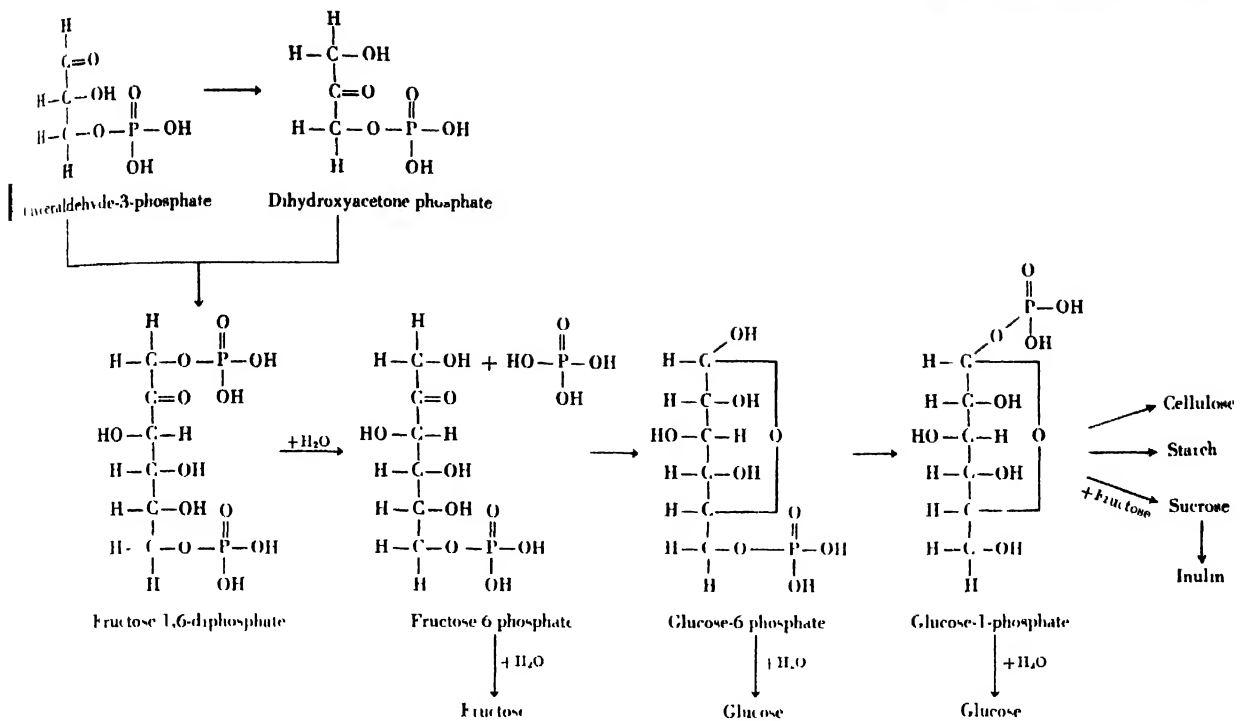
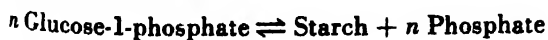


Fig. 5. Formation of carbohydrates from glyceraldehyde-3-phosphate in plants.

Significant quantities of glucose may be metabolized or transformed by another route, the pentose phosphate pathway. In this pathway, glucose-6-phosphate is oxidized to 6-phosphogluconic acid which then undergoes oxidative decarboxylation to form ribulose-5-phosphate. The further transformations of ribulose-5-phosphate are shown in the diagram (Fig. 7). Enzymes catalyzing these transformations have been obtained from plant tissue extracts. Evidence that these enzymes function *in vivo* is provided by the fact that all of the compounds shown here are normal plant constituents and that  $C^{14}$ -labeled intermediates of the pentose phosphate cycle are quickly converted into other compounds in living tissues. In very young plant tissues the Embden-Meyerhof-Parnas pathway appears to be the only operative mechanism for the dissimilation of glucose. In older tissues there is an increasing development of the pentose phosphate pathway. See CARBOHYDRATE METABOLISM.

**Starch.** This most abundant reserve carbohydrate of the plant kingdom is formed reversibly by the action of the enzyme starch phosphorylase on glucose-1-phosphate



The formation of the linear polymer of starch, amylose, requires the phosphorylase enzyme which is specific for the formation of  $\alpha$ -1,4-glycosidic linkages (Fig. 8). The branched polymer, amylopectin, is formed by the action of a branching enzyme (Q enzyme) which transfers short chains of the linear polymer to the 6-hydroxyl position of the glycosyl residues of an amylose (or amylopectin) chain. In addition to being degraded by starch

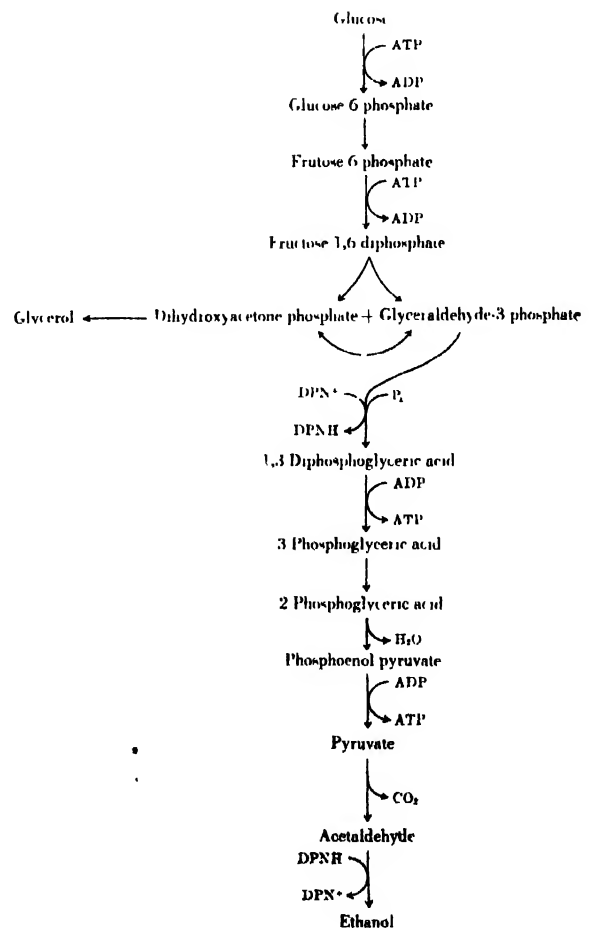


Fig. 6. The pathway of glucose dissimilation in glycolysis.

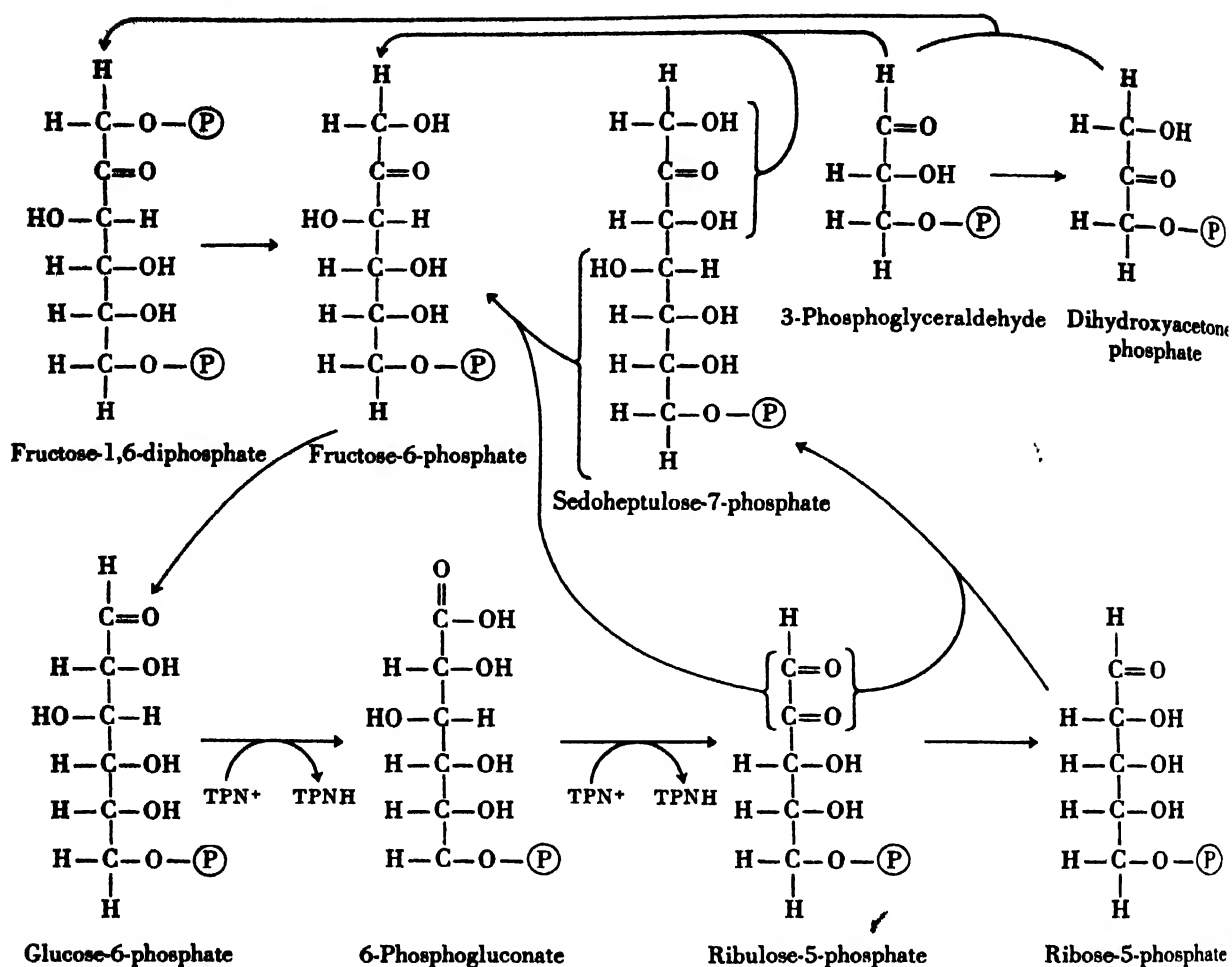
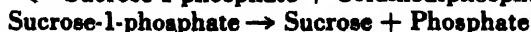
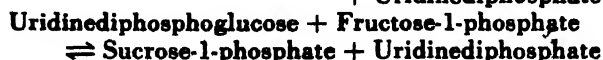
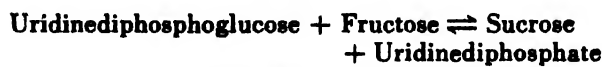


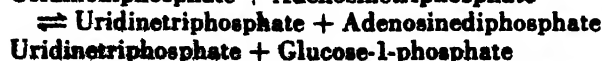
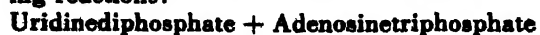
Fig. 7. Pentose phosphate pathway.

phosphorylase, starch is hydrolyzed to maltose by the combined action of  $\alpha$ - and  $\beta$ -amylases. The  $\alpha$ -amylases are specific for the hydrolysis of the 1,4-glycosidic linkages in the interior portions of the starch molecules and consequently rapidly lower the viscosity of a starch suspension, producing dextrins as end products. The  $\beta$ -amylase attacks starch molecules at the nonreducing ends, cleaving off maltose units one at a time. In seeds which contain a high percentage of starch, there is a severalfold increase in amylase activities during germination. See STARCH.

**Sucrose.** In higher plants sucrose synthesis occurs by the following reactions:



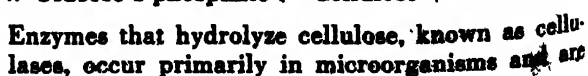
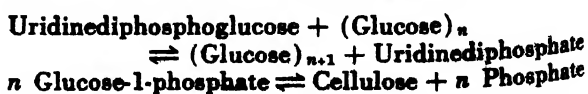
Uridinediphosphoglucose is formed by the following reactions:



The conversion of sucrose to glucose and fructose is by a hydrolytic cleavage catalyzed by the enzyme sucrose. See SUCROSE.

**Pentoses and pentosans.** Ribulose-5-phosphate and xylulose-5-phosphate occur as intermediates in the photosynthetic cycle, whereas ribulose-5-phosphate and ribose-5-phosphate are intermediates in the pentose phosphate cycle. Ribose and xylulose can arise by the hydrolysis of their phosphate esters. Xylulose is isomerized to xylose, which further transforms into arabinose through the intermediate formation of uridine diphosphoxylose and uridine diphosphoarabinose. However, there is no evidence that these free pentoses can serve as precursors for the formation of pentosans. Isotopic-labeling experiments indicate that the pentose residues of pentosans arise from hexoses by loss of carbon atom 6. See CARBOHYDRATE.

**Cellulose.** It appears that cellulose is synthesized by one or both of the following reactions:



Enzymes that hydrolyze cellulose, known as cellulases, occur primarily in microorganisms and are

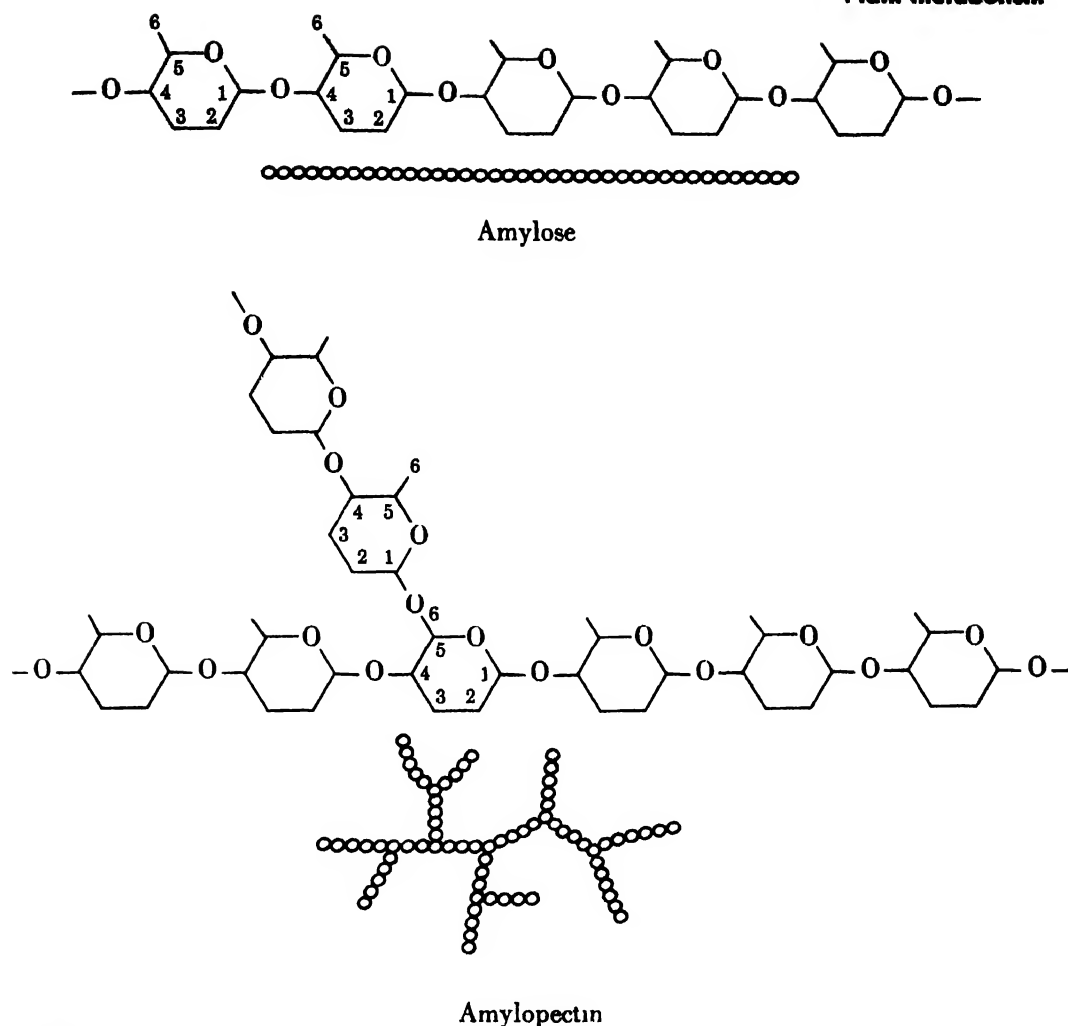


Fig 8 Linear and branched polymers of starch.

responsible for the rapid decay and solubilization of dead plant residues. *See CELLULOSE.*

**Pectin.** The galacturonic acid residues of the pectic compounds appear to arise directly from the glucose skeleton by oxidation of carbon atom 6, and inversion of the configuration of carbon atom 4. The methyl ester group comes from methionine by transmethylation. Enzymes bringing about the degradation of pectic compounds are of two kinds. One, pectin methyl esterase, causes the hydrolysis of the ester linkages to yield pectic acid and methanol. The other, pectin depolymerase, cleaves the polymer to galacturonic acid and digalacturonic acid. These enzymes occur widely in plant tissues and in microorganisms (*see MICROBIOLOGY*). Noteworthy is their importance in the ripening of fruit where the characteristic progressive decrease in firmness is the result of degradation of the pectin compounds in the cell structures. *See FRUIT (BOTANY); PECTIN.*

**Ascorbic acid.** This substance is closely related to the carbohydrates and is of nearly universal occurrence in plants but as yet has no clearly established function. Its biosynthesis is as shown in Fig. 9. *See ASCORBIC ACID.*

#### ORGANIC ACIDS

By virtue of their wide distribution in relatively large quantities in the plant kingdom, several aliphatic acids are referred to collectively as the plant acids. These include the acids in the form of ions and esters, for example, malate, citrate, isocitrate, tartrate, succinate, oxalate, fumarate, *cis*-aconitate, oxalacetate,  $\alpha$ -ketoglutarate, and glycolate. These organic acids and their salts constitute buffering systems which undoubtedly are of importance in controlling the pH (acidity) of the cells. Organic acids neutralize the effect of an excessive uptake of cations. For example, when potassium nitrate is taken up by a plant and the nitrate reduced to ammonium ion, there is an excess of cations in the cell. This is balanced by the synthesis of organic acid anions. Thus the organic acid content of a plant grown on nitrate as a source of nitrogen is several times that of a plant grown with an ammonium salt as a source of nitrogen.

**Functions of organic acids.** Several of the plant acids occur as intermediates in the oxidation of acetyl coenzyme A by the citric acid cycle. Because many kinds of organisms operate the citric acid



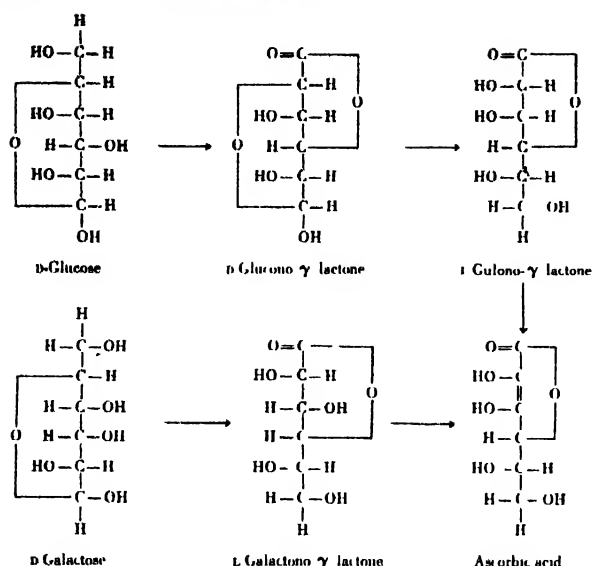
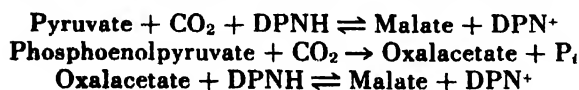


Fig. 9. Biosynthesis of ascorbic acid.

cycle without the accumulation of large quantities of the intermediates, it must be concluded that the accumulations that do occur, such as malate in apples and citrate in tomatoes, are incidental to the operation of the oxidative cycle. These accumulations may result from a block at one point in the cycle or from excessive cations present. All the acids listed above, with the exception of oxalate and tartrate, are in dynamic equilibrium with other plant constituents; that is, the pools of accumulated acids are undergoing metabolic turnover. For example, a single administration of  $\text{C}^{14}$ -labeled glucose to a tomato results in a rapid labeling of the citric acid without any net increase in the total citrate present. Then with time the label disappears from the citrate with no decrease in total citrate present.

Some of the organic acids which are intermediates in the citric acid cycle are also synthesized in significant quantities by other pathways. Malate synthesis occurs by the following reactions, in which  $\text{P}_i$  represents inorganic orthophosphate:



Isocitrate can be formed by a reversal of the isocitric dehydrogenase reaction:



Isocitrate can then give rise to *cis*-aconitate and citrate.

Oxalacetate and  $\alpha$ -ketoglutarate can be produced by the deamination of the amino acids aspartic acid and glutamic acid, respectively.

Tartaric acid and its salts accumulate in large quantities in grapes and are found in smaller quantities in the leaves of many plant species. The rate of synthesis and accumulation of tartrates is slow and further metabolism is extremely slow. The immediate precursors of tartaric acid are not known.

It is clear, however, that the closely related malic acid is not a precursor. Oxalate salts also accumulate in large quantities in some plants through oxidation of glyoxalate. Oxalate is metabolized further to carbon dioxide, formate, or both, at a very slow rate by most plant cells. The table gives an incomplete but representative listing of the occurrence of various organic acids in plants.

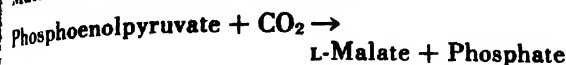
**Acid metabolism of the succulents.** In some species of plants there is not simply a slow accumulation of organic acids but rather a large diurnal fluctuation in the total acids present. These fluctuations are characteristic of a group of plants known as succulents, which are characterized by thickened spongy leaves and stems. Typical succulents occur in the families *Cactaceae*, *Begoniaceae*, *Compositae*, and *Crassulaceae*; the genera *Bryophyllum* and *Sedum* of the *Crassulaceae* have been most extensively studied. The main features of the acid metabolism of succulents are (1) a rapid increase (one- to fivefold) in total acidity at night and an equally rapid decrease in daylight, (2) increase in the rate of acid formation brought about by low temperatures; (3) same intensity of light required to bring about a decrease in total acids as that required for photosynthesis; (4) a very low respiratory quotient ( $\text{CO}_2$  evolved/ $\text{O}_2$  absorbed), less than 0.1, during the first hours of darkness; and (5) a concomitant disappearance of carbohydrates, particularly starch, as the acids appear. The low, and sometimes negative, respiratory quotient values suggested to early experimenters that carbon dioxide was a reactant in the formation of the acids (see PLANT RESPIRATION). Two kinds of evidence confirm this suggestion. First, the rate of acidification of succulent leaves in the dark is a function of the carbon dioxide concentration over the range 0–1%. Second,  $\text{C}^{14}$ -labeled carbon dioxide is incorporated immediately and principally into the malic acid formed. By virtue of the earlier experiments in which the concentration of acids and carbohydrates was measured and of the later experiments with  $\text{C}^{14}$ -labeled carbon dioxide,

Organic acid content of various plant tissues

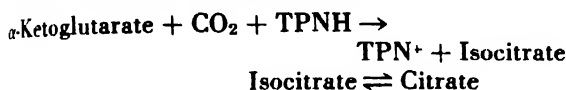
Tissue	Citric	Malic	Iso-citric	Ox-alic	Tar-taric	Suc-cinic
Leaf (meq/100 g dry leaf)*						
Rhubarb	231	22		143		0.5
Bryophyllum	42	166	213	6		4
Nicotiana tabacum	23	68		35		0.06
Tomato	74	61		45		
Spinach	10	9		309		
Valencia orange	27	68		87		
Fruit (meq/100 ml juice)						
Valencia orange						
Immature	62	2				
Mature	27	2.6				
Lemon						
Immature	73	4.4				
Mature	106	2.9				
Apple juice	3–15	0.3				

\* meq = milliequivalents.

acids, and carbohydrates, the following details of acid metabolism of succulents are established. Malate is formed by a carboxylation reaction



and is the chief participant in the diurnal fluctuation of acid content. The phosphoenolpyruvate arises primarily from carbohydrates. In the light, malate is converted by way of the hexoses into starch. In the dark, starch is converted by way of the hexoses (and pentoses) to phosphoenolpyruvate which is then carboxylated by carbon dioxide produced in respiration. Citrate participates to some extent in the diurnal fluctuation of acidity. It may be formed by a carboxylation reaction

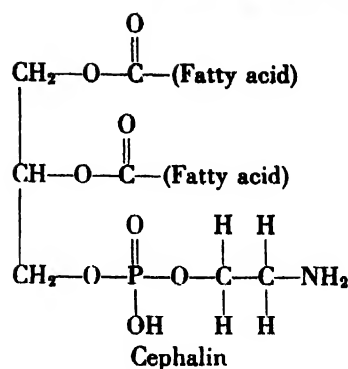
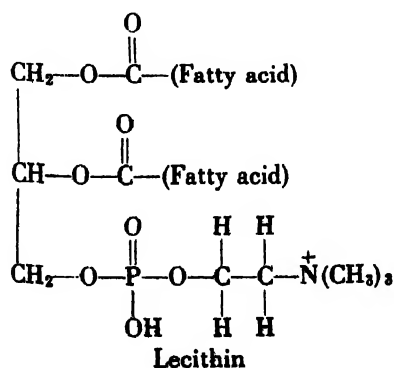
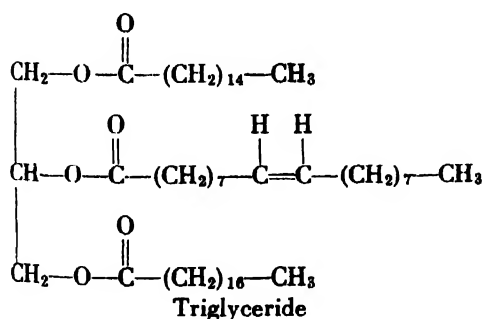


or by the conversion of malate to citrate.

Although isocitric acid is a major constituent of succulent plants, it is not subject to diurnal variations in concentration. It accumulates slowly in the leaves and after the addition of  $\text{C}^{14}$ -labeled carbon dioxide, in either light or dark, becomes labeled very slowly, in contrast to malate and citrate.

#### LIPID METABOLISM

The lipids include a group of substances that yield fatty acids on hydrolysis. These are the fats and oils (triglycerides), in which the fatty acids are esterified with glycerol; the waxes, in which the fatty acids are esterified with long-chain monohydric alcohols; and the phospholipids, in which the fatty acids are esterified with glycerol or with other alcohols with which phosphoric acid and basic nitrogenous constituents are combined.



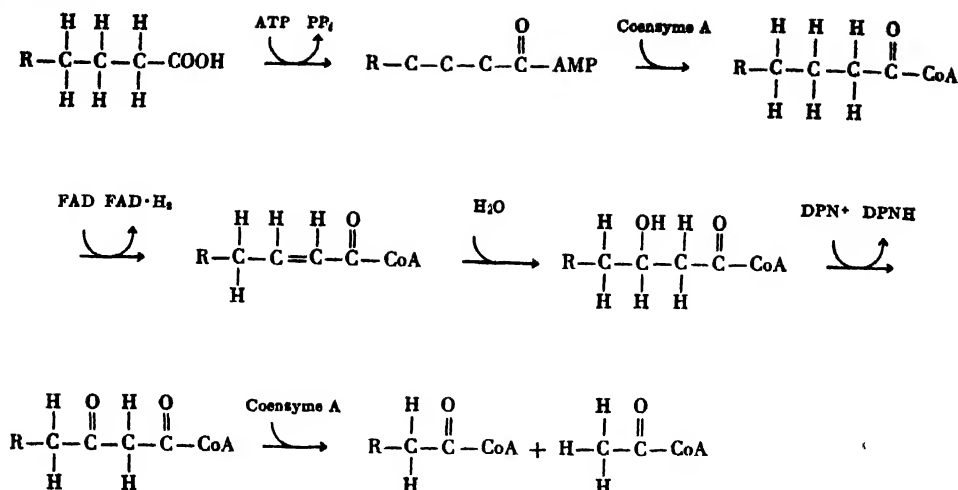
#### Lipids

The constituent fatty acids of the lipids may be saturated or unsaturated and may contain 4–26 carbon atoms. With few exceptions these acids are unbranched and contain an even number of carbon atoms. The leaf fats contain linolenic, linoleic, and oleic acids as principal fatty acids. Fruit fats, such as palm, avocado, and olive, contain principally palmitic and oleic acids, whereas there is a great variation in the fatty acid composition of seed fats. The structure of the triglycerides that have been characterized suggests that there is complete randomization in the formation of triglycerides from fatty acids and glycerol; that is, fats and oils consist of mixtures of mixed triglycerides rather than of mixtures of pure triglycerides.

The fatty acids are synthesized by the condensation of acetate units (acetyl CoA), followed by reduction. Triglycerides are formed by the condensation of two molecules of fatty acid acyl coenzyme A with  $\alpha$ -glycerophosphate, dephosphorylation of the phosphatidic acid formed, followed by the condensation of a fatty acid acyl CoA with the diglyceride. Lecithins are formed by the transfer of the phosphorylcholine moiety from cytidine diphosphocholine to a diglyceride. Cephalins are formed by an analogous reaction utilizing cytidine diphosphoethanolamine.

The degradation of triglycerides occurs by way of a lipase-catalyzed hydrolysis to glycerol and fatty acids. The fatty acids are converted to acetyl CoA as shown in Fig. 10.

There are two periods in the life cycle of certain higher plants when fat metabolism dominates other biochemical events. In the oil-bearing species there is a brief period during the formation of the seed in which the oil content of the seed increases ten- to thirtyfold. During this period the respiratory quotient of the seed increases to values as high as 1.5, indicating that the triglyceride synthesis occurs in the seed itself from a more highly oxidized substrate, such as carbohydrate. During the germination of oil-bearing seeds there is a rapid conversion of triglycerides to carbohydrates which are then translocated to the growing portions of the seedling. This conversion is manifest by low respiratory quotients (down to 0.3). In germinating seeds the rapid metabolism of fats is accompanied by the synthesis of enzymes that are required for this rapid metabolism. See LIPID METABOLISM.

Fig. 10.  $\beta$ -oxidation of fatty acids.

## PRINCIPAL FATTY ACIDS OF PLANTS

Acid	Formula
Lauric	$\text{CH}_3-(\text{CH}_2)_{10}-\text{COOH}$
Myristic	$\text{CH}_3-(\text{CH}_2)_{12}-\text{COOH}$
Palmitic	$\text{CH}_3-(\text{CH}_2)_{14}-\text{COOH}$
Stearic	$\text{CH}_3-(\text{CH}_2)_{16}-\text{COOH}$
Arachidic	$\text{CH}_3-(\text{CH}_2)_{18}-\text{COOH}$
Palmitoleic	$\text{CH}_3-(\text{CH}_2)_5-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-(\text{CH}_2)_7-\text{COOH}$
Oleic	$\text{CH}_3-(\text{CH}_2)_7-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-(\text{CH}_2)_7-\text{COOH}$
Linoleic	$\text{CH}_3-(\text{CH}_2)_4-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-\text{CH}_2-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-(\text{CH}_2)_7-\text{COOH}$
Linolenic	$\text{CH}_3-\text{CH}_2-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-\text{CH}_2-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-\text{CH}_2-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-(\text{CH}_2)_7-\text{COOH}$
Ricinoleic	$\text{CH}_3-(\text{CH}_2)_5-\overset{\text{OH}}{\underset{ }{\text{C}}}(\text{H})-\text{CH}_2-\overset{\text{H}}{\underset{ }{\text{C}}}=\overset{\text{H}}{\underset{ }{\text{C}}}-(\text{CH}_2)_7-\text{COOH}$

## NITROGEN METABOLISM

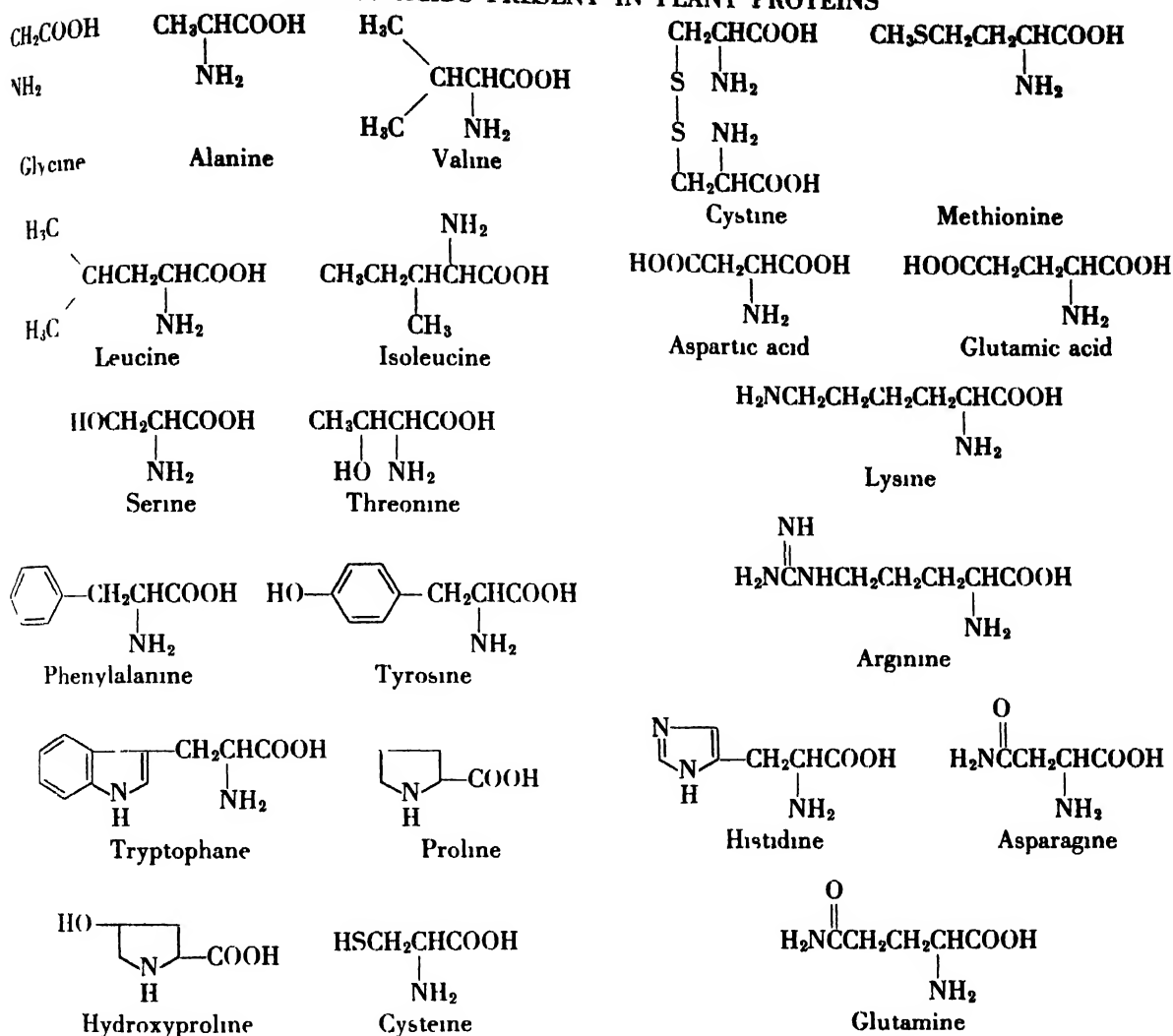
**Amino acids, amides, and proteins.** The principal nitrogenous constituents of plants are proteins. Although there are many kinds of protein in plants, they, like all proteins, are made up of a relatively small number of different kinds of amino acids condensed together through the elimination of a molecule of water between the carboxyl group of one amino acid and the amino group of another. With two exceptions, proline and hydroxyproline, all of the amino acids which have been obtained from protein hydrolyzates are  $\alpha$ -amino acids. The  $\alpha$ -carbon atom is a center of asymmetry so that two optically active isomeric forms are possible for each amino acid. Only the L-amino acids have been found in protein molecules. In addition to the protein amino acids, many other amino acids occur

free or in simple organic compounds. Although some D-amino acids are known in biological materials, the different kinds and quantities are much more limited than for the L-amino acids.

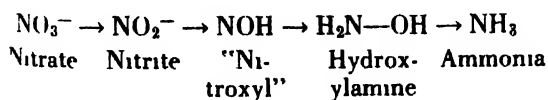
The carbon, hydrogen, and oxygen for the formation of amino acids in plants comes from carbon dioxide and water, by way of photosynthesis.

Soil nitrate is the principal source of nitrogen for higher plants (see NITROGEN CYCLE). The occurrence of nitrate in soil is the result of the microbial nitrification of the ammonia produced by the degradation of plant and animal residues. Nitrate is absorbed into the roots of plants and in most species is translocated as such to the aerial portions of the plant, where it is reduced to ammonia. Root cells also are capable of reducing nitrate to ammonia and in some plants a major fraction of the nitrogen translocated out of the

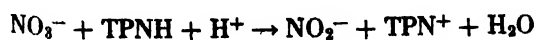
## AMINO ACIDS PRESENT IN PLANT PROTEINS



roots is in the form of amino nitrogen. The pathway of reduction of nitrate is through a series of two electron transfers.



Nitrate reductase, the enzyme catalyzing the first step, has been obtained from *Neurospora* and from the tissues of many higher plants. It is a molybdo-flavoprotein which transfers electrons from TPNH to nitrate.

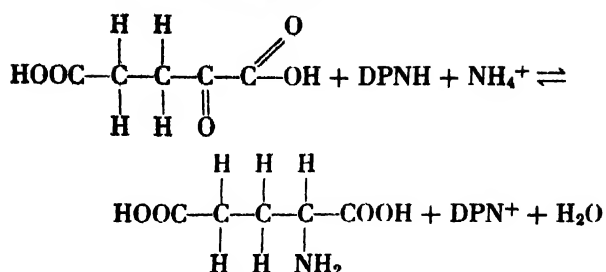


The discovery that the enzyme requires molybdate anion as an activator provides a biochemical explanation for the requirement of plants for molybdate. Although nitrite reductase and hydroxylamine reductase activities have been observed in plant extracts, these enzymes require further characterization with respect to coenzyme and metal ion requirements.

Significant quantities of nitrogen gas from the air are reduced to ammonia by free-living bacteria in the soil and by the *Rhizobia* which grow in symbiosis with the leguminous plants. See LEGUME; SOIL MICROBIOLOGY.

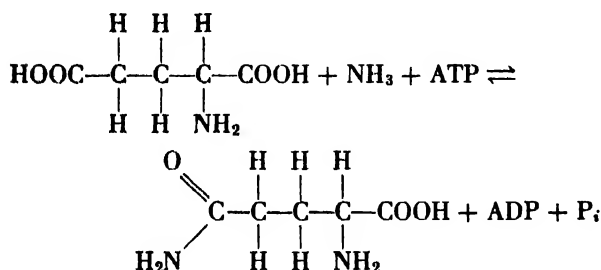
Ammonia is incorporated into the organic nitrogen compounds of plants through the amination of  $\alpha$ -ketoglutarate, oxalacetate, and pyruvate, to form glutamic acid, aspartic acid, and alanine, respectively. Experiments with  $\text{N}^{15}$ -labeled ammonia have shown that the amino acids and proteins of plants undergo constant breakdown and resynthesis. In such experiments glutamate and aspartate are the first amino acids to become labeled with  $\text{N}^{15}$ . The principal reaction for incorporating ammonia into amino groups appears to be that catalyzed by glutamic dehydrogenase.

The amino acids formed by direct amination of  $\alpha$ -keto acids can transfer their amino groups by transamination reactions to other  $\alpha$ -keto acids to form the corresponding amino acids. There are many different transaminase enzymes, all of which require pyridoxal phosphate as a coenzyme. Pro-



line and hydroxyproline cannot be formed by transamination reactions. Proline is formed from glutamic acid by way of glutamic acid semialdehyde and pyrroline carboxylic acid. Hydroxyproline is formed by oxidation of proline. Higher plants are able to synthesize all the amino acids which they utilize in their metabolism. Some of the lower plants and some bacteria are not able to synthesize all the amino acids which they require and are therefore dependent upon plant and animal residues to supply these preformed.

Although ammonia is an intermediate in the metabolism of living cells, ammonium salts are extremely toxic if allowed to accumulate in the cell. Living organisms prevent the accumulation of ammonia by excreting it or by combining it into some innocuous compound. Higher plants incorporate ammonia into the amides glutamine and asparagine. Glutamine is synthesized by the following reaction:



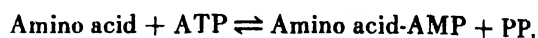
The exact manner in which asparagine is formed is not known. *See ASPARAGINE.*

In seedlings, especially etiolated seedlings, a dominant feature of nitrogen metabolism is the formation of glutamine, asparagine, or both. This synthesis of amides is a result of the high level of ammonia produced during the enzymatic hydrolysis of the seed proteins and the utilization of the amino acids to supply energy and carbon skeletons for the synthesis of new cells in the seedling. As the seedlings develop and the level of ammonia drops, the amides gradually disappear as their nitrogen is utilized in the synthesis of other amino acids to be used in protein synthesis.

The prominent feature of the nitrogen metabolism of any growing cell is the synthesis of the cellular proteins. Because there may be as many as several thousand different kinds of proteins in a cell, it is obvious that protein synthesis in a rapidly growing cell is a result of very complex processes. Proteins may differ from one another in amino acid composition, in amino acid sequence, and in the three-dimensional folding of the long polypeptide chains. These characteristics are, of course, under

genetic control (*see GENETICS*). Some of the most interesting developments in biology in this century center around the discovery that the genetic material of cells consists of nucleic acids and of nucleoproteins, and that the biosynthesis of proteins and nucleic acids is interdependent and simultaneous.

Proteins appear to be synthesized directly from amino acids, by an energy-dependent (ATP) series of reactions. The first step in this series of reactions is the activation of the amino acids. There is probably a separate activating enzyme for each amino acid:



in which  $\text{PP}_i$  indicates inorganic pyrophosphate. In subsequent steps the activated amino acid is transferred to an acceptor or carrier which further transfers the activated amino acid into the polypeptide chain that is to become the finished protein molecule. Although it is not yet possible to state exactly the point of interdependence of this series of reactions with nucleic acid synthesis, it is under investigation in many laboratories throughout the world. The proteins which are enzymatically active are presumably synthesized by the same pathway as other proteins. However, special interest attaches to the synthesis of enzymes because of the interesting problem of the amino acid sequence and the configuration of the polypeptide chain at the active site of the protein.

In most tissues the bulk of the protein consists of the enzymatically active proteins which catalyze the reactions characteristic of living cells. In seeds, however, the bulk of the proteins present appears to serve the function of storage proteins to be utilized as a source of amino acids by the young seedling before it becomes established in the soil. *See SEED (BOTANY)*. These reserve proteins are hydrolyzed by the action of proteolytic enzymes present in the seeds. In many seeds there is a manifold increase in this proteolytic activity during germination (*see PLANT GROWTH; REPRODUCTION. PLANT*). As mentioned earlier, this rapid degradation of proteins to amino acids leads to the production of large quantities of ammonia which become incorporated into glutamine and asparagine. *See AMIDE, ACID; AMINO ACIDS.*

An interesting observation relates to the protein metabolism of excised leaves. The total protein content of a mature leaf remains constant or declines slowly as the leaf ages. After the excision of the leaf, there is a rapid decline of the protein level. This decline is checked by the induction of adventitious roots on the leaf. *See LEAF (BOTANY)*. These experiments suggest that the protein metabolism of leaves is partially under the control of some factor or factors produced by root cells. *See ROOT (BOTANY)*. Such observations are of general interest because of the possible connection with the normal decline and senescence of mature leaves and fruit.

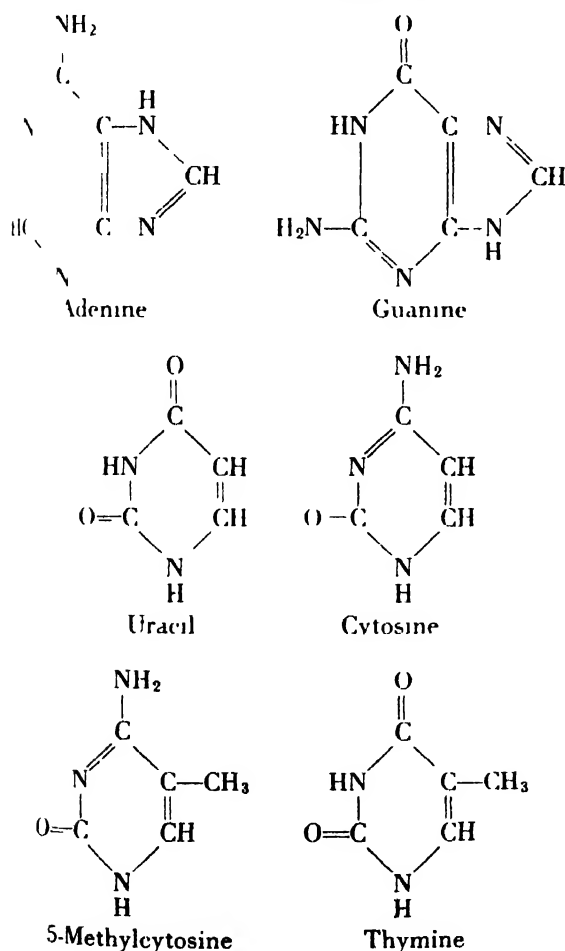
The infection of and multiplication of virus in plants is a special case of protein synthesis of intrinsic interest. In tobacco leaves infected with tobacco

[tobacco mosaic virus there is a rapid synthesis of virus protein. This is a *de novo* synthesis from free amino acids rather than from preexisting leaf proteins. Once the virus protein is formed it does not undergo turnover; that is, there is no degradation and resynthesis of virus proteins. See PLANT VIRUS.

**Purines, pyrimidines, and nucleic acids.** Purines and pyrimidines are of importance because of their universal occurrence in living cells as components of coenzymes and nucleic acids.

The purines, adenine and guanine, are well-known constituents of nucleic acids and of nucleotide cofactors, such as adenosinetriphosphate, adenosinediphosphate, DPN, TPN, CoA guanosinetriphosphate, and guanosinediphosphate. Other purines (hypoxanthine, xanthine, and uric acid)

#### PURINES AND PYRIMIDINES



found in plants have no established function. The pyrimidines (uracil, cytosine, 5-methylcytosine, and thymine) also occur as nucleotide cofactors, for example, uridinediphosphate, uridinetriphosphate, and uridinediphosphoglucose, as well as in nucleic acids. The biogenesis of the purine ring utilizes nitrogen from the amide group of glutamine, from glycine, and from aspartic acid, and carbon from glycine, formate, and carbon dioxide. Pyrimidine ring biogenesis also results from the utilization of simple precursors (aspartate, carbon

dioxide, and ammonia) in a series of enzymatic reactions.

The occurrence of the purine and pyrimidine bases as nucleotide units in nucleic acids is of great importance because the nucleic acids as such, or combined with proteins to form nucleoproteins, are the carriers of genetic information. Nucleic acids are high-molecular-weight substances containing many nucleotide units linked as shown.

Purine-ribose-phosphate

Pyrimidine-ribose-phosphate

Ribonucleic acids (RNA) contain ribose and are characteristically found in the cell cytoplasm, whereas deoxyribonucleic acids (DNA) contain deoxyribose and are found only in the nucleus (see CELL NUCLEUS; CYTOPLASM). The nucleotides of adenine, guanine, cytosine, and uracil are found in RNA, whereas those of adenine, guanine, cytosine, thymine, and 5-methylcytosine are found in DNA. Examination of the hydrolysis products of DNA has shown that the molar ratio of adenine to thymine is unity, and that the molar ratio of guanine to cytosine + 5-methylcytosine is unity. This apparent pairing of the purine and pyrimidine nucleotides, along with the x-ray diffraction patterns of DNA, has led to the important concept that the nucleotide chains of DNA are coiled in a double helical structure. The sequence of purine and pyrimidine bases along the DNA molecule are thought to be of significance in relation to the genetic properties of the molecule. Mitotic division clearly requires some mechanism for the exact replication of those DNA molecules which carry hereditary traits (see MITOSIS). The synthesis of both DNA and RNA is intimately linked to protein synthesis. The determination of the exact manner of these syntheses and their interrelation is a major problem of biological research. See HETEROCYCLIC COMPOUNDS; NUCLEIC ACID.

**Alkaloids.** Of the alkaloids or nitrogenous bases which have a widespread occurrence in the plant kingdom, only the purines and pyrimidines have recognized functions. The others are thought to represent end products of nitrogen metabolism. Historically, man's knowledge of and interest in the alkaloids stems from their physiological effect on animals. Nicotine and nornicotine of tobacco; atropine of *Atropa*, *Datura*, and *Hyoscyamus*; the ergot alkaloids from the fungus *Claviceps purpurea*; and reserpine from *Rauwolfia* are examples of alkaloids that have been useful as pharmacological agents. In the tobacco plant nicotine and nornicotine (Fig. 11) are synthesized by roots and transported to leaves. Ornithine is the precursor of the pyrrolidine ring of nicotine. The pyridine ring is derived from nicotinic acid. The methyl group may come from methionine or from the  $\beta$ -carbon of serine. Atropine also appears to be synthesized in the roots of the plants in which it occurs although it accumu-



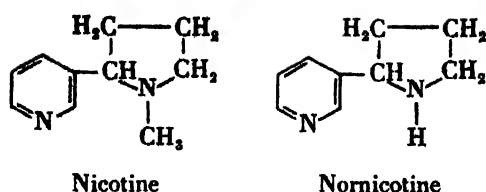
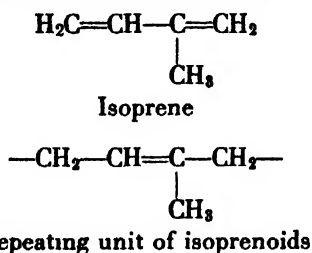


Fig. 11. Structural formulas of nicotine and nornicotine.

lates in the leaves. Very little is known of the biogenesis of the many other alkaloids. The determination of the manner of their synthesis, of the conditions under which they arise, and of their possible function in plants would be a significant contribution to plant metabolism. See ALKALOID.

#### TERPENE METABOLISM

**Oils, resins, and carotenoids.** The terpenes and their derivatives may all be considered as being derived from isoprene. This group of substances includes essential oils, resins, the carotenoid pigments, phytol, and rubber. Probably all plants are able to synthesize some of the carotenoid pigments



and phytol (the  $C_{20}$  alcohol present in chlorophyll). The ability to produce significant quantities of essential oils and resins, however, is much more restricted in the plant kingdom, perhaps to 2000 of 400,000 species. The oils and resins are produced by specialized cells or groups of cells and may either accumulate in these cells or be secreted into resin ducts. See SECRETORY STRUCTURES, PLANT.

Carotenoids, deposited as plastid pigments, occur in roots, stems, leaves, flowers, and fruit. In green leaves, the carotenoids are associated with chlorophyll in the chloroplasts (see CAROTENOID; CHLOROPHYLL). The carotenoids are synthesized in the tissue in which they are found and transport does not seem to occur. Acetate, as acetyl CoA, appears to be the precursor of all the isoprenoid compounds. The intermediate steps probably involve first the formation of acetoacetyl CoA, then the formation of the 6-carbon compound  $\beta$ -methyl- $\beta$ -hydroxy glutaryl CoA which is converted into a 5-carbon compound that can react with itself to form compounds containing 10, 15, 20, 30, or 40 carbon atoms. See FAT AND OIL, EDIBLE; FAT AND OIL, NONEDIBLE; RESIN.

**Rubber.** Rubber occurs in hundreds of species of plants and may constitute up to 20% of the dry weight of the plant. It occurs as small particles suspended in an aqueous serum. This suspension, or latex, occurs in and is formed in specialized

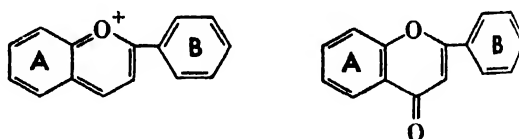
cells or vessels. The latex which is obtained by tapping the *Hevea* tree contains all of the enzymatic apparatus necessary for incorporating radioactive acetate into rubber. The repeating unit in the chemical structure of rubber is isoprene.

Although the deposit of large quantities of rubber in the latex vessels of plants represents a possible source of energy for the plant, there is no evidence that it does actually function as a reserve food. See RUBBER.

#### OTHER METABOLITES

**Tannins.** The tannins are a group of astringent substances apparently formed by the condensation of phloroglucinol and its derivatives, or by the condensation of gallic acid, *m*-digallic acid, ellagic acid, caffeic acid, and quinic acid. Although the tannins have become useful items of commerce for example, in leather tanning, little is known of their biogenesis, and their function in plant metabolism is unknown. See TANNIN.

**Anthocyanins and flavones.** The water-soluble pigments of plants consist almost entirely of anthocyanins and flavones. These pigments occur largely as glycosides and yield, on hydrolysis, a sugar and a 2-phenylbenzopyrylium derivative (anthocyanidin), or a sugar and a 2-phenylbenzopyrone (flavone) derivative. A great variety of anthocyanidins exist. These differ from one another by the degree of substitution on the  $\beta$ -ring, by the total number of hydroxyl groups, by the degree of methylation of these hydroxyl groups, and by the number (1 and 2) and kinds of sugar residues linked to the anthocyanidin by the glycosidic linkage. Similarly a wide variety of flavone pigments occurs in plants. These pigments (Fig. 12) are generally produced



2-Phenylbenzopyrylium anion  
(anthocyanidin)

2-Phenylbenzopyrone  
(flavone)

Fig. 12. Structural formulas of 2-phenylbenzopyrylium anion and 2-phenylbenzopyrone.

by plants under high light intensity and low nitrogen and phosphorus supply. In some cases however, pigment production is independent of light, and may occur in roots or in etiolated seedlings. Although the chemistry of these pigments is understood, little is known of their biogenesis, and nothing of their metabolic function. See ANTHOCYANIN; FLAVONE; GLUCOSIDE; see also PLANT PHYSIOLOGY.

[J.E.V.]

**Bibliography:** J. Bonner, *Plant Biochemistry* 1950; M. Dixon and E. C. Webb, *Enzymes*, 1958; J. S. Fruton and S. Simmonds, *General Biochemistry*, 2d ed., 1958; O. H. Gaebler (ed.), *Enzymes Units of Biological Structure and Function*, 1956.

W. D. McElroy and B. Glass (eds.), *Inorganic Nitrogen Metabolism*, 1956; R. Robinson, *The Structural Relations of Natural Products*, Weizmann Memorial Lectures, 1, 1955; G. C. Webster, *Nitrogen Metabolism in Plants*, 1959.

## Plant morphogenesis

A biological science concerned with the origin and development of the form and structure of plants. It deals with those phenomena, both internal and external, by which form is determined. Other names sometimes given to morphogenesis are experimental morphology, causal morphology, experimental embryology, and *Entwicklungsmechanik*.

Form is the result of the orderly growth of various dimensions. The form of the body and of such structures as leaves, flowers, hairs, and cells are examples of this controlled development. Form and structure in living things are visible expressions of the organizing character of life at every level.

**Correlation.** This is a general term for the relationship among the parts of an organism, especially during its development. Most bodily parts grow at about the same rate; thus, form does not alter greatly during development. In some cases, however, one part grows faster than other parts, or one dimension faster than another. Even here the relationship among growth rates remains fairly constant so that form changes in a regular and predictable fashion. Sometimes the physical basis of correlation is evident. For example, a terminal bud may inhibit or greatly retard the growth of buds below it.

**Polarity.** In almost all organisms an axis appears during development and its two ends become different, as root and shoot in higher plants. This may be seen very early, even in the fertilized egg itself. It is evident in regeneration. Polarity also appears in physiological processes. In many cases, there is a descending gradient in rate of metabolic processes from one end of the axis to the other. Movement of substances may also be polar. It has been suggested that bioelectrical factors are involved in polarity. In some cases rearrangement of materials by centrifugation has been found to change the polar axis, and it may be reversed by other means.

**Symmetry.** The arrangement of lateral structures around the main polar axis is not irregular, but shows patterns of symmetry of various sorts. Most plant axes are vertical, and the distribution of leaves around them (phyllotaxy) is radial and precise. The points of attachment of leaves may be opposite, whorled, or dispersed in a spiral (alternate). Such spirals fall into definite classes which have simple arithmetical relationships to one another.

Where the axis is horizontal, the upper and lower sides are usually different (dorsiventrality). The structures on the right and left sides are usually mirror images of each other (bilateral symmetry). Organic symmetry is far less precise than that of crystals, to which it has sometimes been compared.

**Differentiation.** A conspicuous fact of development is that during its course the parts of the organism become different from one another. An example of such differentiation is that between the two ends of the polar axis. Development in higher plants is accompanied by the progressive appearance of distinct organs—roots, stems, leaves, flowers, fruits, and seeds—the outward sign of division of labor within the plant. Growth and differentiation generally proceed together but there are cases where one occurs without the other. In the course of the life cycle of an organism, differential changes occur in a definite series of stages.

Internal differences that arise among tissues and among cells during development are as marked as is external form. Sometimes such a difference can be traced to a particular cell division in which the two daughter cells are unlike. In more primitive organisms differentiation proceeds as a series of such divisions, but in most cases the precise point of origin of these differences is hard to locate. Particular patterns of internal structural differences have an hereditary basis, but are often modified by factors in the internal or external environment. Even within a single cell there is often a considerable degree of differentiation among its parts. The fact that cells usually have the same number of chromosomes as the fertilized egg (or a simple multiple of these) suggests that the genetic constitution of every cell is like that of all others. Why differences should arise among them, therefore, seems to be due either to different exposure to external factors or to changes in the cytoplasm of the cells.

Physiological differences between parts of the plant are common. Cells of the root are unable to synthesize sugar and certain other necessary substances and must get them from the shoots. Many localized physiological differences are maintained because of the impermeability of the cell membranes to substances produced in the cells.

**Regeneration.** The ability of many organisms, especially during their early developmental stages, to replace lost parts is a distinctive phenomenon of morphogenesis. Cuttings from the stems of many plants will produce roots if placed in a favorable environment.

Though this potency becomes progressively restricted in animals as the individual matures, it may remain almost indefinitely in plants. In many cases mature plant cells which have been induced by chemical means to divide again will then give rise to small growing points from which a whole plant may be formed. Some plants naturally reproduce themselves by small plantlets produced from marginal leaf cells.

**Atypical growth.** The organizing control under which a plant develops may sometimes break down, resulting in the production of galls, tumors, cancers, and other abnormal forms. Of particular morphogenetic interest are galls formed on plants in reaction to certain insect stimulation, especially from the cynipid wasps. These galls are usually of a very

specific shape and internal structure, which vary with the particular insect and the plant species. Substances introduced by the insect or formed from the larva hatching from its egg evidently have a very definite morphogenetic effect. Certain rather formless plant galls, the "crown galls," produced by the attack of various bacteria, have some resemblance to animal cancers.

The most extreme case of disorganization occurs in tissue cultures, where single cells or formless masses of tissue continue to grow without developing into an organism.

**Ecological factors.** Various factors, in the organism itself or in its environment, have been found to be important in the determination of its form and structure. Since most plants are stationary, they are in general more susceptible to environmental influences than are animals, which can move from unfavorable to favorable surroundings. Some of the more important external and internal factors are discussed in the following paragraphs.

**Water.** This is of special importance in plants. In relatively dry soil, the cuticle of a particular species tends to be heavier and the tissues more woody and made up of smaller cells. This is now thought to be a direct effect of water shortage on differentiation. In plants which can grow with their shoots either in air or submersed in water, the differences in development and structure under the two environments are very marked.

**Temperature.** An important effect of temperature on plants has been found in the early growth of the seedling. If sprouting seeds of grains and some other plants are exposed to relatively low temperatures, the early developmental stages are passed through rapidly. Seeds thus treated, when placed under normal conditions, produce the final mature state of the plant much earlier than untreated ones. By this process, called vernalization, winter wheat will grow to maturity as soon as spring wheat.

**Light.** This factor is of particular morphogenetic importance in plants. Vegetative development is affected by intensity of light, low intensities in some cases producing spindly or etiolated growth. The quality of light is also influential. The red end of the spectrum tends in many cases to stimulate growth of reproductive structures and the blue end, differentiation and vegetative growth.

Important in the production of floral organs is the duration of exposure to light, or the photoperiod. Where this is relatively long in relation to the dark interval, flowering is stimulated in "long-day" plants. In "short-day" species, however, flowering occurs only when the daily photoperiod is relatively short and alternates with a long uninterrupted dark period. The factors involved in the transformation of the vegetative to the flowering state are not all well understood, but the production and transfer of specific substances which influence differentiation are probably involved.

**Mechanical factors.** Stresses and strains of various sorts have some importance in development. Gravity, through its effect on the distribution of substances affecting form, influences various geotropic orientations in plants. A tree swayed by air currents tends to have its widest diameter in the plane of sway. A tree guyed firmly by cables will not grow in diameter as fast as one that is unsupported.

**Chemical substances.** Chemical factors of many kinds affect development. Sometimes the result is very specific and radical, as when a teasel plant, grown in soil of high nitrate content, develops a much twisted stalk instead of the straight one formed in ordinary soil. The relative amount of substances may also be important. Thus plants in which the ratio of carbohydrates to nitrogen is high will tend to produce flowers and fruits, but such plants will usually form only vegetative organs when this ratio is low.

**Hormones and growth substances.** Certain more complex chemical substances, especially the "chemical messengers" or hormones, profoundly affect development in plants. Various substances have been found in plants which control development. Most notable is auxin, indoleacetic acid, which has various effects. It stimulates or influences cell wall elongation, particularly in the early stages of development, and thus plays a role in most of the bendings or tropisms of plants made in response to gravity and light. It stimulates root development, cambial activity, and the growth of seedless fruits. It may inhibit the formation of the corky (abscission) layer that causes leaves and fruits to drop off. Auxin in a terminal bud prevents or retards the growth of buds below it for a certain distance from the apex, and thus affects the branch system of the plant. It is involved in the formation of crown-gall tumors. Many of these effects may also be produced by various synthetic substances, chiefly other organic acids.

**Grafts and chimeras.** In plant chimeras the tissues of two individuals may be intimately combined, either naturally or artificially. The two outer cell layers of a plant may be derived from tomato, for example, and the remainder from nightshade, or vice versa. These are related but quite dissimilar species, and the contribution of each to the form of leaf, flower, and fruit may thus be determined. No genetic union is produced, and the sex cells formed by the chimera will be identical with those of the species that provides the layer of cells just under the epidermis.

**Hereditary factors.** Many single genes have been found that control the form of the body or of its parts. The shape of leaves, flowers, and fruits in many plants are examples of form that are gene-controlled. Such genes evidently control the relative rates of growth in various directions. Where form changes with size, genes for size will evidently have an indirect effect on form. Genes that check growth at particular embryological stages will have

most pronounced effects on those structures that are most rapidly developing at those stages. They may disturb the correlative effects between structures or alter hormone distribution.

Chromosomes influence development apart from the genes they carry. In members of a polyploid series in plants, for example, the size of the organism, its parts, and its cells is often proportional to the number of chromosome sets, from  $4n$  down to  $1n$ . The larger the number of sets, also, the shorter the polar diameter of leaves and fruits tends to be as compared to the equatorial diameter. See PHOTOPERIODISM IN PLANTS; PLANT GROWTH; PLANT HORMONES. [E.W.S.]

**Bibliography:** J. T. Bonner, *Morphogenesis*, 1952; K. Goebel, *Einleitung in die experimentelle Morphologie der Pflanzen*, 1908; D. W. Thompson, *On Growth and Form*, 2 vols., 1952; C. W. Wardlaw, *Morphogenesis in Plants*, 1952.

## Plant movements

Movements of attached members of nonmotile plants are divided into two categories: those which are responses to external stimuli and those which result from stimuli of internal origin. If responses to external stimuli are mediated by growth, they are known either as tropic or nastic movements.

In the tropic responses the direction of the curvature or movement is dependent on the orientation of the stimulus. The fundamental feature of tropic stimulation is the space distribution of the energy which is applied. The tropic curvatures are brought about by unequal elongation (growth) rates on opposite sides of the plant organ, which, in turn, are dependent upon the unequal distribution of growth substances called auxins (see AUXIN). Tropic curvatures are either toward (positive tropism) or away from (negative tropism) the source of stimulation.

In contrast to the tropisms, growth movements of plants in which the plant component response is determined by the internal structure but without reference to the orientation of the stimulation energy, are defined as nastic movements. Responses of attached members of nonmotile plants to stimuli of internal origin may be dependent or independent of growth. Elongation, tropisms, nutations, and certain nastic responses are dependent on growth, whereas various specialized pulvinal movements (discussed later under turgor movements) are mediated by mechanisms other than growth. See LEAF (BOTANY).

In addition, movements of motile organisms or free plant parts in response to external stimuli are known as tactic movements. These are distinguished from movements of similar organisms in response to stimuli of internal origin, which are sometimes termed autonomic locomotions.

Plant movements will be discussed in the following order: phototropism, geotropism, electrotropism, thigmotropism, haptotropism, traumatot-

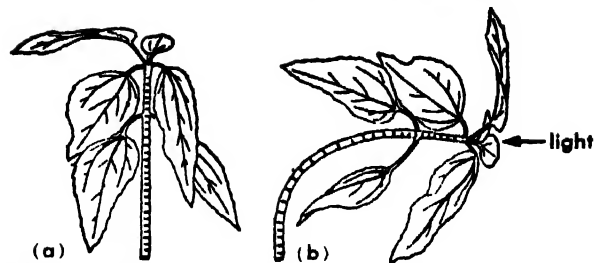


Fig. 1. Positive phototropism. (a) Apex of a sunflower plant with its stem marked at equal intervals. (b) Same plant after having been illuminated on one side. Bending occurs in the region of elongation. (From G. M. Smith et al., *A Textbook of General Botany*, 5th ed., Macmillan, 1953)

ropism, hydrotropism, chemotropism, nastic movements, tactic movements, nutation, and turgor movements.

**Phototropism.** The bending of plant organs mediated by unequal growth rates of the opposite sides and directed toward or away from the source of visible light is called phototropism (Fig. 1). Ordinarily, stems and other aerial organs are positively phototropic, whereas roots and similar structures respond negatively. See ROOT (BOTANY); STEM (BOTANY). This generalization, however, has a number of exceptions. It is known that some stems are negatively phototropic; some roots are positively phototropic; others are nonphototropic. Some structures, such as the *Avena* coleoptile (a widely used material for the study of tropisms), are either positively or negatively phototropic, depending upon the amount of incident energy. Low dosages of light (100 meter-candle-seconds) induce positive bending, which is known as the first positive curvature. Much higher dosages of light may cause either positive or negative tropism. However, little is known about negative phototropic curvature of stems.

In *Avena* coleoptiles and certain other stems and roots, the extreme apex is the most sensitive to light. Removal of the tip results in a sudden loss or diminution of phototropic sensitivity. Because primarily the apical cells are concerned with the production of auxin, there is apparently a close relationship between auxin production and phototropic sensitivity.

Unilateral light falling on the sensitive organ is absorbed by a pigment, and a light gradient is established across the organ. The photoreceptor is a yellow pigment, but its specific identification has not been made. The best evidence indicates that either riboflavin or carotene absorbs the radiant energy, which mediates the first positive curvature (see CAROTENOID; RIBOFLAVIN). After the radiant energy has been absorbed, it must in some way mediate an excess of auxin on the shaded side of the coleoptile. It seems that physical continuity between the illuminated and shaded side is essential for maximum curvature to develop, and that

some means of amplification is necessary to account for the differences in auxin concentration. Three possible mechanisms could be involved in the attainment of the asymmetry of auxin distribution.

The first mechanism requires the more effective photodestruction of auxin on the lighted side. According to the original observation made by F. W. Went, dosages of light which induced the first positive phototropic curvature caused the total auxin production to be lowered from 100 to 84%. (Photodestruction of auxin *in vitro* also has been demonstrated.) The auxin content of the lighted side apparently is lowered to 27% (compared with 50% in a dark control), whereas in the shaded side the content apparently increased to 57%. In other experiments the ratio has been reported to be as low as 11 to 89. Thus, although it is apparent that photodestruction of auxin may be involved in phototropic curvature to some extent, differential destruction of auxin by light presumably does not function as an amplification mechanism. It still remains important to explain how lateral transport of auxin can be effected.

The transport of auxin against a concentration gradient requires an oriented force. Because auxin is an acid which will migrate toward the positive pole in an electrical field, the possibility of such functional electrical polarity has been proposed. It is known that unilateral illumination by white light causes the extreme apex of the *Avena* coleoptile to establish a transverse electrical polarity such that the lighted side becomes electronegative to the shaded side. Such polarity would cause the auxin to be transported to the positive and shaded side. It has been shown that transversely applied direct current can induce lateral transport of auxin. In positive phototropism this viewpoint is further supported by the fact that externally applied electrical fields can increase or decrease light-induced bending, depending upon the polarity of the electrical stimulation. Lateral transport of auxin, caused by the photoinduced transverse electrical polarities, remains as a possible, but not the only, explanation of positive phototropic curvature.

It is reasonably well established that visible light alters auxin synthesis in various plants. The extreme apex of the coleoptile is the site of auxin synthesis. Thus, it is possible that the unequal distribution of auxin is brought about by an asymmetric auxin synthesis. All that is required is more effective destruction or inhibition of an enzyme or cofactor required for the synthesis of auxin in the illuminated side (see ENZYME). This would allow the auxin precursor to accumulate on the lighted side to such an extent that it would rapidly diffuse to the shaded side where it could be converted to active auxin. Such a sequence would decrease the auxin concentration on the illuminated side and increase the auxin concentration on the shaded side, resulting in positive curvature. The evidence which supports this concept also is indirect.

Because roots normally grow beneath the surface of the soil, there is very little functional signifi-

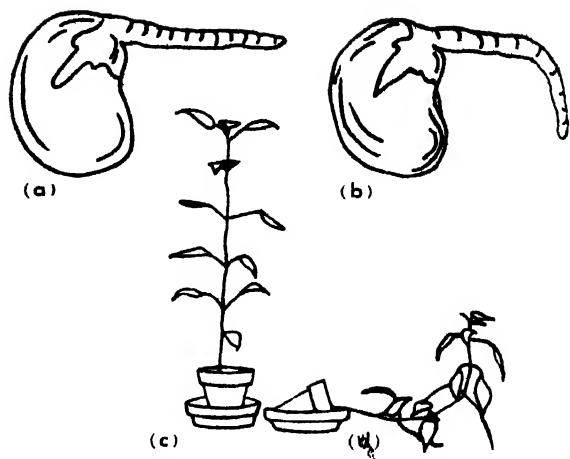


Fig. 2. Positive and negative geotropism. (a) A bean seedling whose primary root has been marked at equal intervals and placed in a horizontal position in a moist chamber. (b) Same seedling 24 hours later. Bending occurs in the region of elongation. (c) A plant of *Iresene* in an upright position. (d) A plant which has been turned on its side. In *Iresene* the negative geotropic response to stimulus of gravity consists principally in upward curvature of the younger nodes (From G. M. Smith et al., *A Textbook of General Botany*, 5th ed., Macmillan, 1953)

cance of the phototropic responses of these organs. It seems reasonably certain that an asymmetrical distribution of auxin is essential for phototropism of roots, but it is not clear how such lateral distribution is brought about. Furthermore, the auxin-to-growth linkage remains relatively obscure.

**Geotropism.** The curvature response of plant components when they are placed in the horizontal position is called geotropism. Primary roots bend toward the force of gravity, hence are positively geotropic (Fig. 2). Shoots or stems bend upward and away from the force of gravity, and thus are negatively geotropic. Geotropic curvatures are brought about by unequal rates of elongation in the upper and lower halves of the organ. As in phototropism, these unequal rates of growth are in turn dependent upon asymmetrical distribution of auxin. Because a direct chemical effect of a change in the direction of the force of gravity is most unlikely, the problem of lateral transport has occupied the major share of attention.

In negative geotropism, most of the available information indicates that transverse movement of auxin is indeed accomplished. A sample of such data is shown in Fig. 3. Only fragmentary evidence contradicts this viewpoint. Elucidation of the mechanism required for the lateral transport of auxin again has paramount significance. Several possibilities have been suggested.

Structures which are negatively geotropic establish a transverse electrical polarity when they are placed in the horizontal position. The lower side becomes electropositive to the upper side. Such polarity is established before the unequal distribution is possible, the polarity appears long before

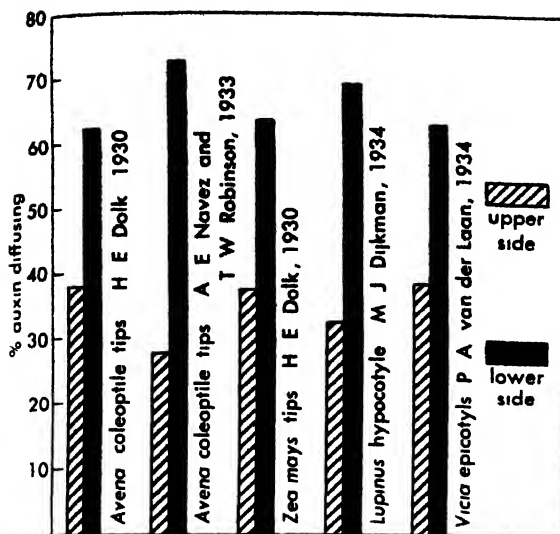


Fig. 3 Comparative percentages of auxin in upper and lower sides of young stem tips.

initiation of upward curvature, and development of the polarity is independent of auxin concentration. Also, these inherent bioelectrical polarities can be reversibly shunted by the use of various dilute solutions of electrolytes. Such treatment also inhibits geotropic curvature.

Another view is that gravity causes the displacement of a statolith, such as a starch granule, which, in turn, by some mechanism which has not been revealed, eventually brings about the auxin translocation. A statolith could conceivably function within individual cells, but it is difficult to understand how it might be involved in the transport of auxin from one cell to the next.

The standard explanation of positive geotropism is based on the premise that total auxin concentration of the roots is above the optimum for growth stimulation. The implication is that any subsequent increase in auxin concentration causes inhibition of growth. Thus, an increase in auxin in the lower half of the root induced by geotropic stimulation inhibits growth of this half. Similarly, a decrease in auxin content of the upper half causes an increase in the rate of elongation. The combination of these changed elongation rates results in downward curvature.

It is generally accepted that auxin moves from the upper to the lower half of the root tip itself when the root is placed in the horizontal position. In this way the supply of auxin to the elongating cells decreases in the upper half and increases in the lower. The mechanism which is responsible for the auxin movement in the tip has not been elucidated. There is limited evidence pointing to the possibility that geotropic stimulation alters the rate of auxin synthesis, and the production of an endogenous inhibitor of growth in the lower half of the root has been suggested. However, the fundamental significance of these processes in positive geotropism has not been extensively evaluated.

Plagiotropic structures, such as lateral stems and secondary roots, have their long axis inclined obliquely away from the vertical axis of the plant. The term diageotropic is applied to such structures as rhizomes when the angle formed with the vertical axis is a right angle. The position maintained by these various organs very likely is a manifestation of a composite response to a number of stimuli, including geotropic stimulation.

**Electrotropism.** This is a growth curvature response to transversely applied direct current or electrical fields. In the *Avena* coleoptile, the initial apical bending is toward the electropositive pole of the stimulating circuit. Figure 4 shows electrotropic curvatures at the indicated times after stimulation when current was applied 5 mm below the apex. In the range of 5–30  $\mu$ a applied for 2 min the magnitude of curvature is dependent upon the current strength. Decapitated and auxin-depleted coleoptiles do not respond to direct current. Additional indirect evidence indicates that electrotropic curvatures of stems are mediated by growth mechanisms which require auxin.

Auxin-depleted coleoptiles which have 3-indoleacetic acid (IAA) applied to the apical ends will respond electrotropically in much the same way as intact seedlings. The magnitude of curvature is directly dependent on IAA concentrations to a maximum of 0.8 mg/liter. It has been demonstrated that electrotropic curvatures are mediated by the asymmetric distribution of IAA which is brought about by lateral transport. A state of polarity in the plant tissue, rather than the flow of externally applied current, is directly responsible for the translocation of the IAA. It is not apparent why the IAA should be transported toward the side that was made electronegative by the applied current.

**Thigmotropism.** A curvature response to mechanical stimulation is called a thigmotropism. When *Avena* coleoptiles or etiolated seedlings or tendrils are stroked or lightly tapped on one side, they respond by bending toward the stimulated side. Because such responses do not occur in coleoptiles depleted of auxin, these curvatures apparently involve differential growth rates caused

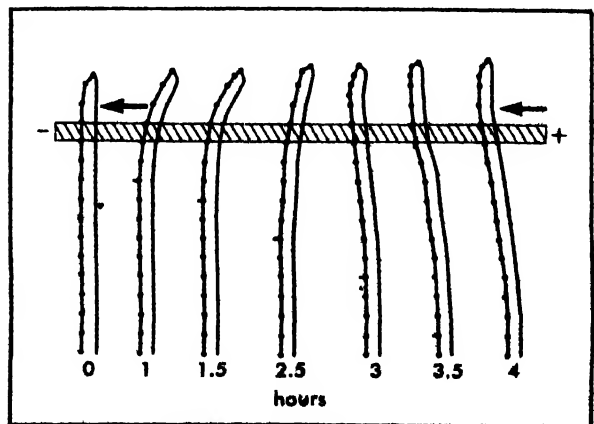


Fig. 4. Electrotropic curvatures of *Avena* coleoptile.



by auxin. It is known that mechanical stimulation can inhibit or prevent negative geotropism and positive phototropism and that such stimulation induces a transverse electrical polarity in the *Avena* coleoptile. The stimulated side becomes electropositive to the opposite side. Thus the assumption is that an asymmetric distribution of auxin is implicated in thigmotropism and that such distribution is mediated by lateral transport to the electronegative side. Neither of these assumptions has been verified. See TAXIS.

**Haptotropism.** Some plants classified as non-climbers have ordinary structures such as the petiole, leaf tip, or specialized tentacles that are sensitive to contact stimulation. It appears that growth of the cells on the contacted side is inhibited, whereas the cells on the opposite side continue to grow. This combination results in a sharp curvature and the plant structures grow around the object which was contacted. Curvatures of this type are defined as haptotropism.

**Traumatotropism.** The curvature of plant organs toward or away from a wound or an injured area is known as traumatotropism. Most curvatures of this type are positive; the direction of bending is toward the wound. These responses in stems and roots are commonly explained as being induced, to a large extent, by interference with the mechanisms for distribution of auxin. Additional factors, which probably are involved to a lesser extent, are the interference of the upward translocation of food factors and the possible destruction of auxin by substances released from the wounded cells.

**Hydrotropism.** The growing of roots toward wetter regions in the soil is called hydrotropism. This type of response can be demonstrated by growing the roots of certain plants in a moist porous material in a shallow tray with a wire mesh bottom. The roots first grow downward and through the wire mesh. They then bend back, away from the dry air below, and toward the moist air. In order for this to happen, the hydrotropic stimulus must overcome the effects of the geotropic stimulus.

**Chemotropism.** The growth response of plant structures to chemical compounds in the environment is defined as chemotropism. The tubes of some pollen grains, when allowed to germinate on nutrient agar in which pieces of pistil have been embedded, will grow toward the pieces of pistil. It is presumed that substances which attract the pollen tubes are diffusing from the pistils.

**Nastic movements.** Curvatures or movements caused by external forces, but whose direction is determined by the internal structure of the responding plant system, are called nastic movements.

**Epinasty.** In epinasty the leaves and stems always tend to curve downward regardless of the orientation of the stimulus. For example, in the stems of *Tradescantia*, when the lateral branches are in the vertical position, auxin is transported only along the dorsal side. When the branches are

in the horizontal position, auxin is also transported along the ventral side. Thus epinastic curvature appears to be due to the action of gravity, which causes the asymmetrical transport of auxin. This response is comparable to geotropism, except that the auxin accumulates on the morphologically determined upper side. The horizontal position of lateral branches is attained as an equilibrium position when the geotropic auxin distribution is equal and opposite to the plagiotropic distribution.

**Nyctinasty.** Nyctinastic movements are those induced by changes in temperature and illumination. This combination of stimuli, both of which are more intense during the daytime, causes some flowers to open in the morning and close again at night. The "sleep" movements of certain leaves are generally included in this category. Most likely such movements of flowers are caused by unequal growth rates on opposite sides of the petal, but the movements of leaves are dependent on changes of turgor (water pressure) in the pulvini (enlarged areas at the base of some kinds of leaves and leaflets). Frequently it is difficult to determine which form of energy is the primary stimulus, but in some instances the individual causative factors can be implicated. For example, before the petals of crocus or tulip are completely expanded (mature) they will open when illuminated and close again when darkened, even when kept at constant temperature. This type of response is defined as photonastic curvature. Similarly, at constant light intensity these petals will manifest thermonasty; that is, they will open in warm air and close in cold air.

**Haptonasty.** The movement of marginal tentacles of the leaves of sundew plants in response to contact by an insect are known as haptonastic curvatures (see INSECTIVOROUS PLANTS). The insect is hereby brought into contact with the smaller central tentacles and ultimately trapped. Similarly, in the Venus' flytrap, contact stimulation causes the leaf blade to fold at about the midrib and to come together to trap the insect. Haptonastic movements are also shown by stamens of some species in the plant families Berberidaceae and Compositae, and by stigmas of the genera *Mimulus* and *Strobilanthes*.

**Tactic movements.** Movements of motile organisms and free parts of nonmotile plants in response to external stimuli are called tactic movements.

**Phototaxis.** These are oriented responses of such plants or plant parts as gametes, spores, algae, and fungi to stimulation by directional illumination (phototactic movements). Generally these forms swim by the use of their flagella and move toward light of low intensity (see ALGAE; FUNGI). This is positive phototaxis. When the light intensity is very high, however, the movement frequently is directed away from the light (negative phototaxis). The plasmodia of slime molds (*Myxomycetes*) move away from light by means of their pseudopodic action. See MYXOMYCETES.

**Chemotaxis.** Chemotactic movements are those in which free plants or plant parts migrate toward (positive) or away from (negative) a specific chemical substance. Chemotactic movements are exhibited by some gametes and such forms as the myxameba of the Acrasiales.

**Nutations.** Rhythmical or periodic movements that are exhibited by shoot apices of plants are called nutations. These autonomic movements are the consequence of internally induced variations of growth rates in different parts of the structure. The simplest nutations are the nodding motions of actively growing apices of such components as the epicotyls of beans. See SEFD (BOTANY). These motions are mediated by alterations in growth rates on opposite sides of the epicotyl. In a slightly more involved nutation known as circumnutation, the plant tip appears to be growing upward in a straight line when observed with the unaided eye. However, slight magnification reveals that the tip actually describes an irregular or spiral path as the apex grows. These movements are also induced by variations in growth rates in different regions of the responding structure. Twining may be viewed as an exaggerated form of circumnutation. Stem tips of twining plants, such as morning-glory and bindweed, are generally long, slender, and without leaves. Because supporting tissue is not well developed, the tips of twining plants frequently bend over to approximate the horizontal position. As in geotropism, auxin apparently accumulates in the lower half of the stem (perhaps partly as a result of geotropic stimulation). More rapid growth in the lower and outer portions of the stem causes an upward spiraling movement of the tip and results in a twisted stem.

**Turgor movements.** A few plants which belong to such families as the Leguminosae have special structures called pulvini located at the base of the leaves and petioles (see LEGUME). When stimulated, the pulvini change their shapes as the result of turgor variation and thereby induce rapid movements of the petioles and leaves. The rate of transmission of the impulse in sensitive plants from the point of stimulation to the pulvinus ranges from about 2 to 15 mm/sec. Some evidence indicates that hormone mechanisms are functional in impulse conduction (Fig. 5).

In *Mimosa* there is a central vascular bundle in each pulvinus of the leaf and the spaces between the surrounding parenchymatous cells are large (see PARENCHYMA). The cell walls of the parenchymous layer are thicker in the upper half than in the lower half of the pulvini. It is known that certain tensions are present in pulvini prior to stimulation. Stimulation of *Mimosa*, the so-called sensitive plant, causes water to escape into the intercellular spaces so that the loss of turgor is greater in the lower half of the pulvinus which, in combination with the tension in the upper half of the pulvinus, results in a change in shape of the pulvinus. This change in pulvinar shape is the cause of the movement of the attached petiole or

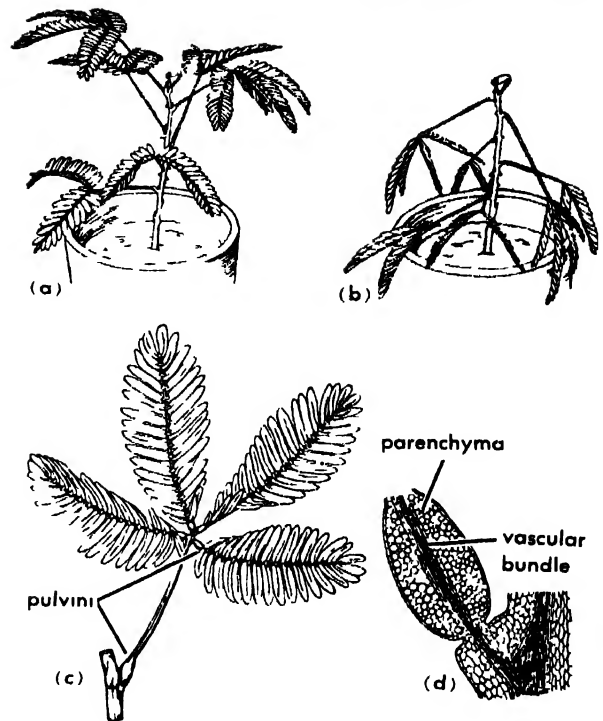


Fig. 5. Turgor movements. (a) *Mimosa pudica*, one of the sensitive plants. (b) Plant of *Mimosa* after leaflets and leaves have responded to a mechanical stimulus. (c) Leaf of *Mimosa* showing the pulvini at base of the petiole and of each primary and secondary leaflet. (d) Lengthwise section of pulvinus of *Mimosa*, diagrammatic. (From G. M. Smith et al., *A Textbook of General Botany*, 5th ed., Macmillan, 1953)

leaf. See PLANT GROWTH; PLANT HORMONE; PLANT PHYSIOLOGY. [A.R.S.C.]

**Bibliography:** P. Boysen-Jensen, *Growth hormones in plants*, 1936; A. C. Leopold, *Auxins and plant growth*, 1955; F. Skoog (ed.), *Plant Growth Substances*, 1951; F. W. Went and K. V. Thimann, *Phytohormones*, 1937.

## Plant names

There are two types of plant names: the common, or vernacular, used in the localities where the plants grow; and the scientific, or Latin, used by botanists universally. Since common names for the same plants often vary in different localities, and since many plants have no common names while others have too many, common names are unsatisfactory for scientific purposes.

Many of the scientific names originated as common names used widely in the days of the ancient Roman Empire, for example, *Salix* (willow), *Quercus* (oak), *Rosa* (rose), and *Viola* (violet). These names were carried over directly into botanical writings of the sixteenth and seventeenth centuries. Some inconvenience was experienced when a single Latin name designated several species of plants, for instance, numerous species of oaks, of roses, or of violets. In such cases it was necessary to add adjectives to designate the different species. Some-

times to indicate a species precisely, several adjectives were necessary and thus the name of the plant became a long and cumbersome descriptive phrase. For greater convenience, later writers attempted to shorten these names. Finally in 1753, Carolus Linnaeus, in his celebrated *Species plantarum*, established the practice of binomial nomenclature—the use of only one descriptive word with the Latin name. This usage proved so convenient that it was at once adopted by botanists. Therefore, the date 1753 is regarded as the beginning of our modern system of taxonomic nomenclature. A binomial is thus composed of the noun, or name of the genus, followed by a word designating the particular species. The generic name is always spelled with a capital letter; most specific words (by some botanists all specific words) are spelled with small letters; for example, *Quercus alba*, white oak.

The generic name, as noted above, in many cases was the exact name of the plant used by the ancient Romans. The vast majority of members of the plant kingdom were, of course, outside the bounds of the ancient world, and, not being known to the Romans, had no Latin names. To give the system universal application, Latinized generic names for such plants had to be invented. Such names as *Fuchsia*, in honor of Leonhart Fuchs, a distinguished botanist, originated in this way.

**Authority for plant names.** It is customary to indicate the name of the botanist who first applied a particular scientific name to a given plant. This botanist is said to be the author and his name (often abbreviated for convenience) is placed after the scientific name of the plant or group; this is the citation. An example is *Trillium erectum* L., a plant named by Carolus Linnaeus. The better known the author, the shorter is the abbreviation of his name. Linnaeus, a name which is very familiar, is abbreviated to L., but few others can be so greatly shortened.

**International botanical congresses.** As the exploration of the earth's surface was extended, thousands of new plants were discovered and given scientific names. In the application of these names, discrepancies in practice became apparent. Personal and national jealousies complicated the situation, and the need for an international accord became evident. The First International Botanical Congress, called by the Swiss botanist Alphonse de Candolle, convened in Paris in 1867. Other congresses were held at Vienna (1905), Brussels (1910), Ithaca, N.Y. (1926), Cambridge, England (1930), Amsterdam (1935), Stockholm (1950), Paris (1954), and Montreal, Canada (1959). These congresses have resulted in the codification of the International Rules of Botanical Nomenclature.

**International rules of botanical nomenclature.** These rules govern the application of scientific names under varying situations. For example, no plant should have more than one scientific name; should more than one, by error, have been applied, the earliest, in general, has priority; all others are synonyms. Likewise, two plants cannot have the

same name. If the same name has been applied to two or more plants, the duplicate names are called homonyms, and the later homonym must be replaced by a new name. Where a plant is first described in one genus and later transferred to another, the original epithet must be retained, unless the same epithet already exists in the new genus. If a conflict such as this develops, a new name must be adopted. These are only a few of a very large number of items covered by the International Rules. In each new congress, the rules become more complicated. See PLANT CLASSIFICATION; PLANT KINGDOM. [E.L.C.]

*Bibliography:* See PLANT TAXONOMY.

## Plant organs

Plant parts having rather distinct form, structure and function. Organs, however, are interrelated through both evolution and development and are similar in many ways.

Roots, stems, and leaves are vegetative, or asexual, plant organs. They do not produce sex cells or play a direct role in sexual reproduction. In many species, nevertheless, these organs, or parts of them (cuttings), may produce new plants asexually (vegetative reproduction). Sex organs are formed during the reproductive stage of plant development. In flowering plants, sex cells are produced in certain floral organs. The flower as a whole is sometimes called an organ, although it is more appropriate to consider it an assemblage of organs. See REPRODUCTION, PLANT.

**Root.** The root is usually the underground part of the plant axis. It may consist of a dominant primary seedling root (taproot) with subordinal branch roots, as in carrots and beets; or it may be composed, as in grasses, of numerous branched roots of similar dimensions (fibrous roots). Collectively, all the roots of a plant are known as the root system. Roots anchor the plant, absorb water and mineral salts in solution from the soil, and conduct these to the stem. Organic food and growth substances received from the stem move to the areas of growth and storage in the roots.

**Stem and leaves.** The stem is usually the aerial part of the plant axis and bears leaves. The stem and leaves together constitute the shoot. In some species the major portion of the stem grows horizontally beneath the surface of the soil, and thus is called a rhizome, or underground stem. The stem conducts water and minerals from the roots to all parts of the shoot, and food materials and growth substances from the shoot to the root. The stem may also serve as a storage organ for water and food. Green leaves containing chlorophyll, when exposed to light and air, carry on photosynthesis. As a by-product of this process, oxygen is returned to the atmosphere. Leaves also return large amounts of water vapor to the air through transpiration (evaporation). Some leaflike structures are protective (bud scales), others are fleshy types in which food and water may accumulate. The first leaves on a seed plant are called cotyledons.

**Flower, fruit, and seed.** The flower is often interpreted as a modified shoot bearing floral organs instead of leaves. These organs are the sepals, petals, stamens, and carpels. The sepals and petals are sterile leaflike appendages. Sepals, like leaves, are commonly green. Petals, containing little or no chlorophyll, are usually white or have some color other than green. The sepals collectively constitute the calyx, the petals the corolla. The calyx and corolla form the perianth or floral envelope. The stamens and carpels are the reproductive floral organs and produce sex cells. A stamen is usually composed of a slender stalk, called the filament, at the tip of which is an anther. The anther is divided into four or fewer pollen sacs in which pollen grains develop. When mature, the pollen sacs open, and the pollen is transferred by wind, water, insects, or man to the tips of the carpels, a process called pollination. The assemblage of stamens is called the androecium, a term implying the male nature of this part of the flower. The carpel assemblage is called the gynoecium to indicate its female nature. Pistil is another term used to designate the female part of the flower. A single carpel may form a pistil, or two or more may be combined into a compound pistil. The pistil generally consists of an enlarged basal part called the ovary. The apex of the ovary usually narrows into a stalk, called the style that terminates in a sticky surface, the stigma. The ovary contains one or more ovules. The egg cell produced in each ovule becomes fertilized by the sperm brought to the ovule by the pollen tube. The latter is an outgrowth of a pollen grain that became attached to and germinated on the stigma. The pollen tube grows through the style to the ovule where it discharges the sperm. After the egg is fertilized by a sperm, the ovary, sometimes together with other floral parts, develops into a fruit. The ovules become seeds.

See separate articles for detailed discussion of the various plant organs. See PLANT PHYSIOLOGY; PLANT TISSUE SYSTEMS. [K.E.]

*Bibliography:* see PLANT ANATOMY.

## Plant physiology

The branch of botany which comprises knowledge of the processes which occur in plants. It is a fundamental tenet of physiology as a science that the usually complex processes occurring in living organisms can be resolved into the relatively simpler processes of physics and chemistry. The field of plant physiology therefore grades imperceptibly into the fields of plant biochemistry and plant biophysics. Much progress in elucidating the detailed mechanisms of the physiological processes that occur in plants has been achieved by using physical and chemical methods as tools of experimentation. No exception has ever been found in which plant processes do not operate in accordance with the fundamental principles of physics and chemistry.

All processes occurring in plants are subject to the dual control of the genetic factors inherent within the plant and the environmental factors to

which it is subjected. The study of the effects of environmental factors upon physiological processes therefore often comes within the purview of plant physiology. In this area of knowledge plant physiology overlaps with the field of plant ecology in a borderline domain of knowledge which is often called physiological ecology.

The effects of genetic factors upon the physiology of plants are under implied consideration whenever the same process is studied comparatively for two different species of plants, or even different varieties of the same species. Differences in physiology between species are as much a reflection of their genetic differences as are their more obvious differences in external morphology. Genetic differences in physiology are often much more subtle than morphological differences; varieties of the same species, indistinguishable morphologically or nearly so, often differ to a marked degree in their physiology.

Investigations of the mechanism whereby specific genetic factors—the genes of the chromosomes—influence physiological processes involves precise probing into the metabolic pathways within cells and into the patterns of enzymatic activity which control such pathways. This subdivision of biology is often called physiological genetics. No sharp boundary can be drawn, however, between this realm of knowledge and the realm of plant physiology.

An intimate relation exists between the cellular structure of a plant and the processes which occur in it. This relationship is a dual one. The organs and tissues of a plant originate as a result of growth, which is itself a complex of coordinated physiological processes. Once materialized, however, the organization of a cell or the cellular structure of a given tissue or organ may have marked effects on the manner in which continuing physiological processes proceed within it. Process and structure are inseparable facets of the phenomenon of growth in plants. See PHOTOSYNTHESIS; PLANT, MINERAL NUTRITION OF; PLANT, WATER RELATIONS OF; PLANT GROWTH; PLANT HORMONES; PLANT METABOLISM; PLANT RESPIRATION. [B.S.M.]

## Plant respiration

A chemical process whereby living protoplasm breaks down certain organic substances with the release of energy which is used in various metabolic activities. In contrast to photosynthesis, in which light (radiant) energy is changed to chemical energy that is bound in organic molecules, respiration consists of a series of transitions in which a part or all of the bound energy is released from organic molecules. For example, sugar is gradually broken down in a series of reactions which involve the release of energy and which may ultimately result in the formation of carbon dioxide ( $\text{CO}_2$ ) and water ( $\text{H}_2\text{O}$ ). See PHOTOSYNTHESIS.

For these reactions to occur, cell catalysts called enzymes must be present to serve a twofold purpose. First, they are instrumental in splitting the

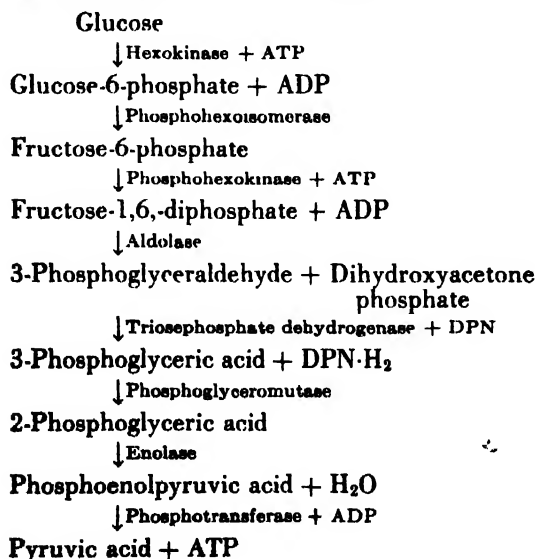
more complex organic molecules into simpler ones. Second, and perhaps more important, these enzymes facilitate the transfer of energy to energy-rich phosphate bonds in organic phosphate molecules. See ENZYME.

**Mechanism of respiration.** For an understanding of the many integrated reactions in respiration, it is first necessary to establish that the living cells of some species of plants have respiratory reactions which do not require utilization of free oxygen. An example of such a species is yeast, in which the following respiratory reaction occurs:



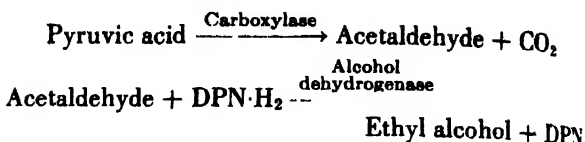
The end products of this anaerobic reaction are ethyl alcohol and carbon dioxide, whereas in the utilization of free oxygen in respiration, carbon dioxide and water are the end products (see PLANT FERMENTATION). Many living cells have both aerobic and anaerobic respiratory systems. Consequently, the kind of respiratory reaction that occurs is generally dependent on whether free oxygen is present. Some species of bacteria cannot live in the presence of free oxygen and are known as obligate anaerobes. In both aerobic and anaerobic respiration, the end products are formed only after a number of intermediate reactions. The intermediate reactions common to both aerobic and anaerobic respiration are called glycolysis and do not utilize or require free oxygen. See BACTERIAL METABOLISM.

**Glycolysis.** The central feature of the initial breakdown by glycolysis is the conversion of hexose sugar to pyruvic acid through a series of reactions which involve phosphorylated derivatives of hexose sugar or other carbohydrates. Each step involves the action of a specific phosphorylase enzyme; these steps may be summarized as follows (enzyme names are shown beside the arrows):



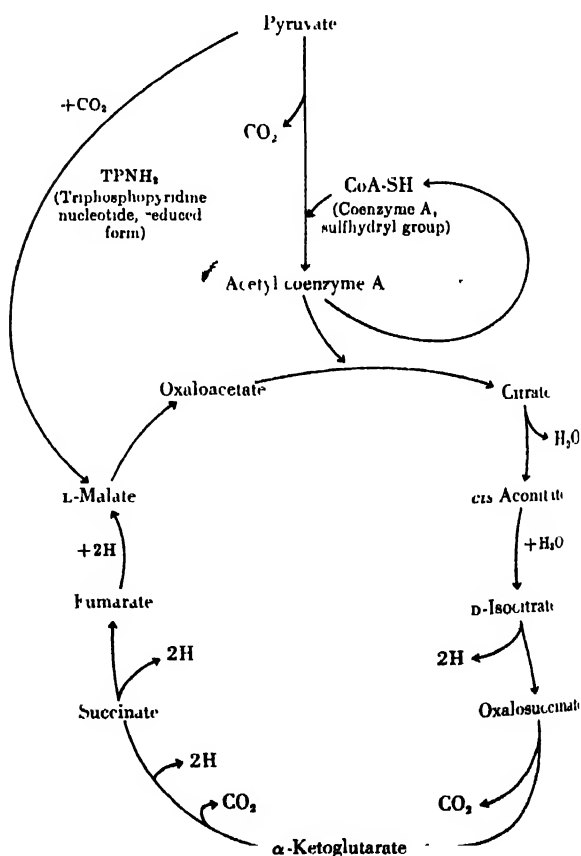
With the formation of pyruvic acid, the initial and intermediate steps common to aerobic and anaerobic respiration end. The fate of pyruvic acid is different in the two kinds of respiration. In one

type of anaerobic respiration the following reactions occur:



**Krebs cycle.** In aerobic respiration pyruvic acid is converted to  $\text{CO}_2 + \text{H}_2\text{O}$  by a series of reactions known collectively as the Krebs cycle. This cycle involves two kinds of enzyme, dehydrogenases which remove hydrogen, and carboxylases which remove  $\text{CO}_2$ . The dehydrogenases, in turn, release the hydrogen to the terminal oxidases, which combine the hydrogen with free oxygen in forming  $\text{H}_2\text{O}$ . Thus the end products of aerobic respiration ( $\text{H}_2\text{O}$  and  $\text{CO}_2$ ) are the final products of pyruvic acid oxidation.

The Krebs cycle may be summarized by the following steps:



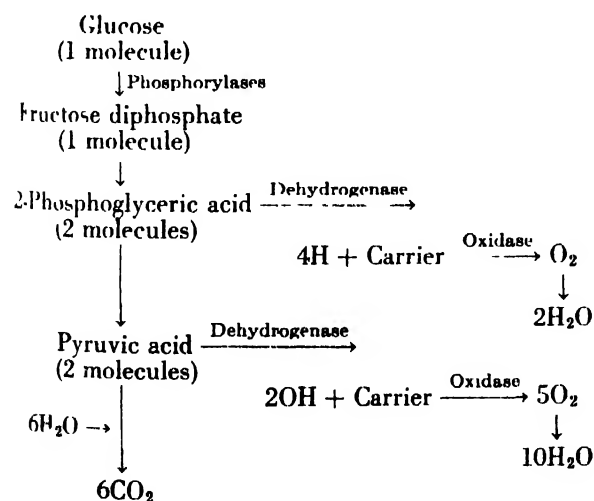
By the addition of the pyruvate (3-carbon compound) to oxaloacetate (4-carbon compound) and by passage through the Krebs cycle, three molecules of  $\text{CO}_2$  are split off from different intermediates and sufficient hydrogen is transferred to oxygen to yield  $\text{H}_2\text{O}$ . In the formation of citric acid (6-carbon compound) from oxaloacetate plus pyruvate, one molecule of  $\text{CO}_2$  is liberated. The citric acid is then degraded via the Krebs cycle with the loss of two additional molecules of  $\text{CO}_2$ . This re-

sults in the regeneration of the 4-carbon compound, oxaloacetate. The cycle is then repeated by adding another molecule of pyruvate. During the cycle, in addition to the  $\text{CO}_2$  formed, some of the intermediate compounds are oxidized by the removal of hydrogen. The hydrogen may react with oxygen by means of certain enzyme systems to form  $\text{H}_2\text{O}$ .

All the molecules of pyruvate that result from glycolysis are not decarboxylated and oxidized to  $\text{CO}_2$  and  $\text{H}_2\text{O}$ . Some of the pyruvate molecules may (1) participate in the formation of high phosphate bond energy enzymes; (2) be intermediate compounds in amino acid synthesis; (3) be precursors, via acetyl coenzyme A, for the synthesis of fatty acids; and (4) be other precursors for the biosyntheses of assimilation products.

Unless some of the pyruvate molecules are utilized in the above-mentioned metabolic pathways, the integrated metabolism that is the basis for life would cease. As far as is known this holds true for most living cells.

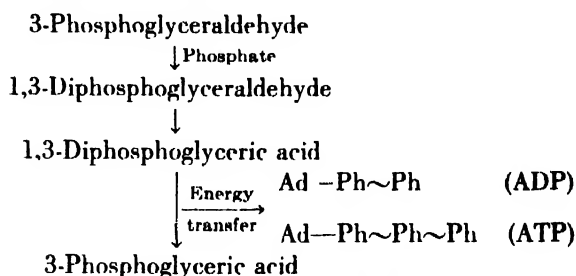
The formation of L-malate from pyruvate by the addition of  $\text{CO}_2$  that is accompanied by the oxidation of  $\text{TPNH}_2$  is known as carbon dioxide fixation or carboxylation. This mechanism is similar to, but distinct from,  $\text{CO}_2$  fixation in photosynthesis. The principal difference appears to be the specific kind of organic molecule that acts as the  $\text{CO}_2$  acceptor. Aerobic respiration can be summarized in its entire reactions as follows:



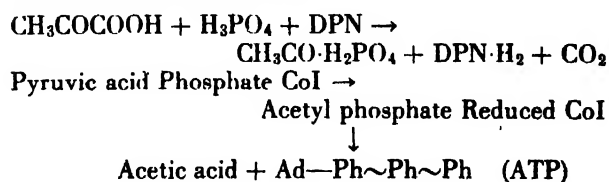
**Energy transfer.** In the above series of reactions in the respiratory process, sugar is broken down by means of a step-by-step oxidation resulting in a loss of energy at each step. Some of this energy is dissipated, for example, as heat energy. More important, however, is the energy which is transferred to the phosphate enzymes and which is subsequently released in various biosyntheses or which is the source of energy for these reactions. This transfer of energy is not shown in the previously discussed reactions because such energy transfer involves additional reactions.

The key phosphate enzymes are the two adenine derivatives. The adenine derivatives are molecules of adenine, ribose, and phosphate groups.

The phosphate groups contain the stored energy. Adenosine is produced from the reaction of ribose and adenine. Adenosinediphosphate (ADP) has two phosphate groups, whereas adenosinetriphosphate (ATP) has three groups. Each phosphate group has 12,000 calories of energy per mole. It is only recently that the mechanism of the adenosine formation has been understood in its relationship to respiration. For example, during glycolysis, ADP and ATP are formed when phosphoglyceraldehyde is converted to 3-phosphoglyceric acid. This may be shown as follows:



The above transfer of energy-rich phosphate bonds (Ph~Ph) can account for only a small portion of the energy which is stored as such bonds. Research has shown that the complete oxidation of 1 mole of hexose to  $\text{CO}_2$  and  $\text{H}_2\text{O}$  should release 680,000 cal. Actual measurements show that approximately 290,000 cal are not released but are stored in the form of phosphate bond energy. It is not entirely clear how the remaining energy is transferred to storage, but it is thought that ATP is generated during the Krebs cycle or pyruvic acid oxidation. Biochemists now are of the opinion that, during the splitting off of hydrogen and carbon dioxide, these reactions involve the taking up of phosphate and the formation of organic phosphate intermediates which, in turn, release the energy-rich phosphate bonds to ATP. The transfer of energy can occur at any of the steps in the Krebs cycle and may be illustrated at the pyruvic to acetic acid step as follows:



A summary of these reactions is given in Table 1 with the numbers of phosphate high-energy bonds yielded. The 16 bonds yield 184,000 cal out of the

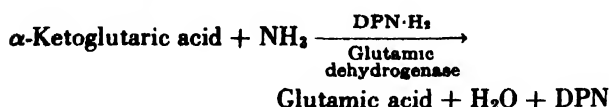
Table 1. Summary of reactions in Krebs citric acid cycle

Reactions	Number of energy-rich bonds
Pyruvic $\rightarrow$ Acetic	4
Isocitric $\rightarrow$ $\alpha$ -Ketoglutaric	3
$\alpha$ -Ketoglutaric $\rightarrow$ Succinic	4
Succinic $\rightarrow$ Fumaric	2
Malic $\rightarrow$ Oxaloacetic	3
Total	16



274,000 cal released by oxidation of pyruvic acid, or an energy capture of 67%.

In the preceding discussion, it has been demonstrated how energy is captured in respiration. Now it becomes important to show how a respiratory intermediate product is synthesized into a cellular constituent. This is the second important function of respiration in terms of metabolism, because  $\text{CO}_2$ ,  $\text{H}_2\text{O}$ , and dissipated energy are useless.  $\alpha$ -Ketoglutaric acid, a respiratory intermediate, is the raw material from which the amino acid, glutamic acid, is synthesized. This synthesis occurs as follows:



**Factors affecting respiration.** From the previous discussion on the nature of the respiratory mechanism, it is evident that respiration involves an extremely complicated series of biochemical reactions which are intimately related to all living processes. If this were not true, methods used for the measurement of respiration would be far more accurate than they are at present. Respiration usually has been measured by the gas-exchange technique. In investigation of the rate of aerobic respiration as it is affected by a certain factor, the oxygen used or carbon dioxide evolved is measured per weight of tissue. The gas-exchange technique thus operates on the assumption that respiration is a complete oxidation of sugar in which one hexose molecule is oxidized by six  $\text{O}_2$  molecules, and six  $\text{CO}_2$  molecules are released. Hence the ratio of the volume of  $\text{CO}_2$  evolved to the volume of  $\text{O}_2$  used is equal to unity. This ratio is called the respiratory quotient ( $\text{CO}_2/\text{O}_2$ ). But as the respiratory mechanism indicates, every molecule of hexose sugar does not always end up as  $\text{CO}_2$  and  $\text{H}_2\text{O}$ . Therefore the rate of respiration as measured by  $\text{CO}_2$  evolution may be far from completely accurate as an index of the respiratory intensity. However,

Table 2. Rates of respiration of various plant species\*

Species	Description	Temperature, °C	Intensity of respiration, ml $\text{O}_2$ / (hr) (g fresh wt)
Cactus ( <i>Cereus</i> )	Herbaceous perennial	12	3.00
Ecuador cholla	Herbaceous perennial	13	6.80
Prickly pear ( <i>Opuntia</i> )	Herbaceous perennial	13	11.40
Stone crop	Herbaceous perennial	13	16.60
Norway spruce	Tree	15	44.10
Common snow-drop	Perennial bulb	13	77.60
Broad bean	Herbaceous annual	12	96.60
Four-o'clock	Herbaceous perennial	15	120.00
Wheat	Annual	13	291.00

\* W. Stiles, *An Introduction to the Principles of Plant Physiology*, Methuen, 1950.

Table 3. Rates of respiration of various plant organs\*

Species	Temperature, °C	Respiratory intensity, ml $\text{CO}_2$ / (hr) (g fresh wt)				
		Sepals	Petals	Stamens	Pistil	Leaves
Mullein	23.0	0.747	0.177	0.761	0.815	0.382
Beard-tongue	23.5	0.571	0.398	0.602	0.689	0.300
Poppy	22.0	0.390	0.367	1.041	0.690	0.332
Tree-mallow	22.0	0.615	0.303	0.576	0.894	0.394

\* From W. Stiles and W. Leach, *Respiration in Plants*, Methuen 1952

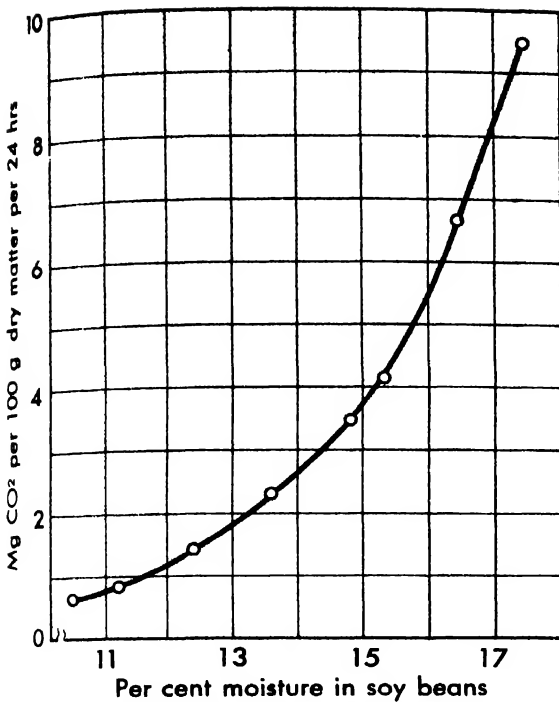
in the development of the present knowledge of factors which affect the rate of respiration, it was necessary to evaluate respiration by gas exchange because better methods were not available to the investigators who obtained the data. Such data do not therefore constitute an absolute indicator of respiratory intensity, but rather indicate an approximation or what might be called an apparent respiration rate.

The rate of respiration is influenced by a complex of interrelated factors; that is, respiration intensity is dependent both on conditions within the living cell and on environmental factors. The role of the living cell in regulating respiratory intensity is illustrated by differences in rate of respiration between species of plants, between organs of the same species, and by age of the organ. Such differences are comparable only if environmental factors are the same. These differences are substantiated by experimental data presented in Tables 2 and 3. These data show that, at approximately equal temperatures, respiratory intensity varies with species of plant and with various plant organs or parts of plant organs of the same species.

**Concentration of substrate.** The discussion turns now to the question of why differences in respiration intensity occur. Excluding environmental factors, the answer is found in the intracellular factors which control the many reactions of respiration. These are concentration of substrate such as sugar, concentration of enzymes, enzyme activators, and enzyme inhibitors; and degree of cell hydration. For example, it has been shown that as sugar substrates are exhausted, the rate of respiration decreases (as measured by  $\text{CO}_2$  evolution). Likewise, decreasing cellular water content in soybean seed results in a reduced rate of respiration (Fig. 1).

**Time factor.** Research on fruit has demonstrated that time is an important factor in respiratory intensity. The time effect as related to age of fruit is illustrated by the data on several varieties of apples (Fig. 2). These data show that as the apples reach maturity, the rate of respiration is increased. Other data show that this peak of respiratory intensity (climacteric) does not continue, but rapidly falls off to a steady state.

**Environmental factors.** The major environmental factors which affect the rate of respiration are temperature, oxygen concentration, carbon dioxide concentration, and light. Respiration is a chemical reaction and thus temperature should increase the



1 The association of the rate of respiration as measured by CO<sub>2</sub> evolution with the moisture content of soybean seed (From P. E. Ramstad and W. F. Geddes *The Respiration and Storage Behavior of Soybeans*, Minn Agr Exp. Sta. Tech. Bull. 156, 1942)

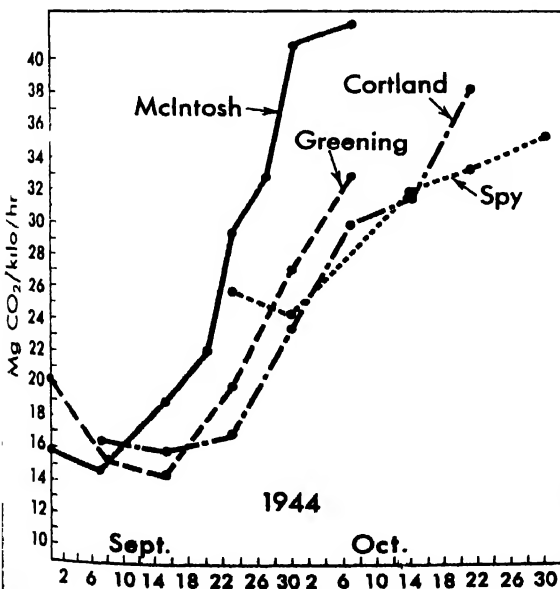


Fig 2 The effect of age of fruit (as denoted by harvest date) on the rate of respiration as measured by CO<sub>2</sub> evolution of several varieties of apples. (From R. M. Smock and C. R. Gross, *Cornell Univ. Agr. Exp. Sta. Mem.* 297, 1950)

rate two to three times for each 10-degree (centide) rise in temperature. This rate of increase of the process is expressed by the symbol  $Q_{10}$  and is called the temperature coefficient. Between 0 and 35°C. the  $Q_{10}$  for most kinds of plant tissue is 2.0. Above 35°C., the rate of respiration can be

shown to increase for short periods of time; but from 40 to 45°C., the rate rapidly falls to less than rates at 30°C. This is thought to be due to enzyme inactivation and failure to maintain a supply of substrate. At 35°C., however, time may affect the respiratory intensity of some kinds of plant organs. The effects of temperature and time on the rate of respiration of plums are graphically presented in Fig. 3.

The considerations relative to factors affecting respiratory intensity are applicable to both aerobic and anaerobic respiration. As previously indicated, the rate of aerobic respiration is affected by the concentration of oxygen as well as by other factors. In closed chambers containing plant material on which respiration is being determined, unless the oxygen used is replaced, the air in the chambers becomes progressively lower in oxygen and higher in carbon dioxide. The rate of respiration is decreased under such conditions. It is difficult to determine from the data whether such a decrease is the result of CO<sub>2</sub> toxicity or lack of O<sub>2</sub>. The direct effect of oxygen concentration can be determined only under experimental conditions where oxygen is the variable and CO<sub>2</sub> is kept at the normal atmospheric level. L. L. Claypool and F. W. Allen studied the effect of oxygen concentration on plum fruit respiration under such experimental conditions, and the data obtained at 86°F are presented in Fig. 4. These data show that with increasing oxygen concentrations from 1 to 50%, accelerated rates of respiration were obtained which maintained a fairly constant rate with time. At oxygen concentrations above 50%, respiratory intensity is accelerated for a short period of time but then falls off.

The direct inhibitory effect of increasing concentrations of carbon dioxide on respiration has been demonstrated by germinating mustard seed. However, increasing CO<sub>2</sub> concentration does not always have a direct inhibiting effect on respiration. Leaves in light placed in higher concentra-

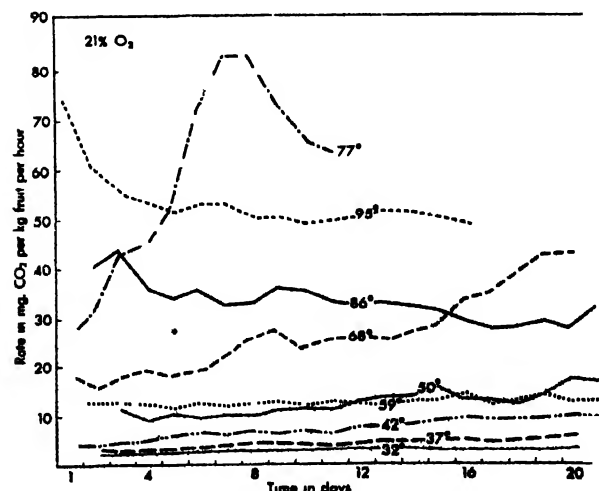


Fig. 3. The effect of temperature on the rate of respiration as measured by CO<sub>2</sub> evolution of Wicksen plums. (From L. L. Claypool and F. W. Allen, *Hilgardia* 21(6):129-160, 1951)

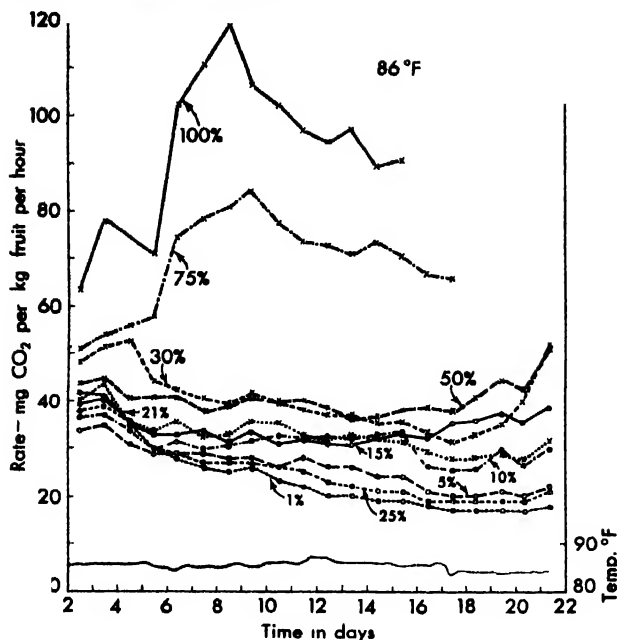


Fig. 4. The effect of oxygen concentration on the rate of respiration as measured by  $\text{CO}_2$  evolution of Wickson plums. (From L. L. Claypool and F. W. Allen, *Hilgardia* 21(6):129-160, 1951)

tions of  $\text{CO}_2$  have higher rates of respiration because the increased photosynthetic rate results in more sugar substrate for respiration.

**Food production and food use.** When food production is compared to food use, the maximum rate of photosynthesis occurs at lower temperatures than the maximum rate of respiration (Fig. 5). This means that during extremely hot daylight hours of summer, the rate at which food is made by photosynthesis may barely equal the rate at which food is used in respiration. Respiration goes on continuously day and night and photosynthesis occurs only during the hours of daylight; thus it is evident that, if such conditions persisted for a very long period, use (respiratory decomposition)

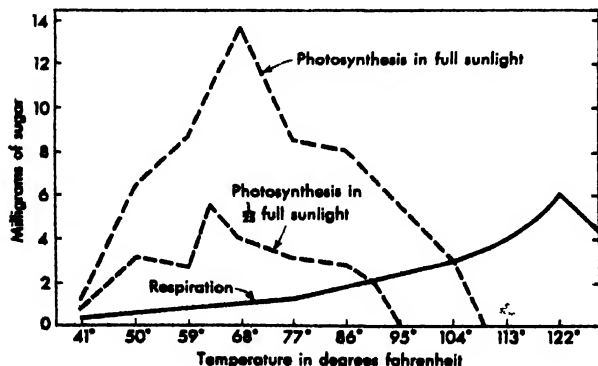


Fig. 5. Relative rates of photosynthesis and respiration in potato leaves during 10-minute exposures to different temperatures in shade and in full sunlight. Recalculated from data by H. S. Lundegardh. (From E. N. Transeau, H. C. Sampson, and L. H. Tiffany, *Textbook of Botany*, Harper, 1953)

would greatly exceed the production of food, growth would cease, and the plant would succumb.

**Respiration and agriculture.** Data from respiratory research have enabled scientists to recommend methods whereby various food products may be transported and stored with a minimum loss of quality. Fruit and vegetables are kept at high quality for shipment and storage by refrigeration. The reason for refrigeration is that low temperature causes respiration rate to be at a minimum. Deterioration in quality occurs at the higher temperatures because of factors associated with high rates of respiration. Deterioration is also held to a minimum by the maintenance of high  $\text{CO}_2$  concentrations and low oxygen concentrations in storage structures, but refrigeration is the most widely used method.

Unlike fruits and vegetables, seeds which are to be stored for food or planting purposes can be kept at high germinability with little deterioration by removal of moisture from the seeds to a level where the rate of respiration is at a minimum. Maintenance of germinability with prevention of loss in quality of stored seeds is now being carried out on a large scale by removal of moisture from seeds to a moisture content which is safe for storage. Moisture removal is accomplished either by natural curing in environments of high evaporation or by artificial drying in humid environments. The safe storage moisture contents of several kinds of seeds are shown in Table 4. A considerable variation exists in safe storage moisture contents among the different kinds of seeds, because safe storage moisture content for a seed is the moisture content which will be in equilibrium with relative humidities of less than 75% so as to prevent or check microbial deterioration. The equilibrium moisture of a particular kind of seed at relative humidities of less than 75% will depend upon the chemical constituents of the seed. Seeds high in hydrophilic colloids (proteins and starches) will have higher safe storage moisture contents than seeds which

Table 4. Safe storage moisture contents of various kinds of seeds for maintenance of high germinability and low deterioration

Kind of seed	Safe storage moisture content, %*
Barley	12.0
Buckwheat	12.5
Blue lupine	11.0†
Corn	12.0
Crimson clover	10.0†
Flaxseed	7.9
Oats	11.8
Peanuts	6.0†
Rice	12.5
Rye	12.2
<i>Sericea lespedeza</i>	8.5†
Sorghum	11.9
Soybeans	9.3
Watermelon	8.4†
Wheat	12.0

\* On a basis of wet weight.

† Determined by Agricultural Experiment Station, Alabama Polytechnic Institute.

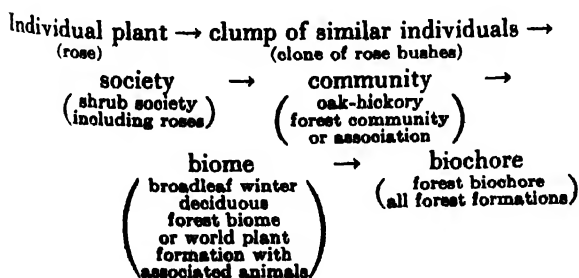
are low in hydrophilic colloids (oil seeds). The same principles are also applicable to the storage of hay crops.

In the discussion of temperature effects on the rate of respiration, the interrelationship between photosynthesis and respiration as influenced by temperature was mentioned. This interrelationship is important in crop yields. The light intensity at which the rates of respiration and photosynthesis are equal is called the compensation point. If light intensities are sufficient to increase photosynthesis rates as rapidly as increases in temperature cause acceleration in respiration, the compensation point changes. However, before growth can occur in plant parts such as stems, roots, bulbs, fruits, and seeds, the rate of photosynthesis must exceed the rate of respiration so that sufficient sugar remains for leaf respiration at night and for translocation to storage organs. Thus two considerations are involved, (1) amount of sugar not used in light, and (2) rate of night respiration, which is generally controlled by night temperatures. A well-known example illustrating these considerations has to do with the size of Irish potato tubers under different conditions. Warm nights, by causing a comparatively high rate of respiration, leave too little sugar for growth of the tubers and small potatoes result. On the other hand, cool nights facilitate yields of large potatoes. Thus in open habitats night temperatures are very important in the growth and development of storage organs. See PLANT GROWTH; PLANT METABOLISM. [J.F.F.]

**Bibliography:** J. Bonner, *Plant Biochemistry*, 1950; J. F. Ferry and H. S. Ward, *Fundamentals of Plant Physiology*, 1959; D. R. Goddard and J. D. Meese, Respiration of higher plants, *Ann. Rev. Plant Physiol.*, 1:207-232, 1950; B. S. Meyer and D. B. Anderson, *Plant Physiology*, 2d ed., 1952; W. Stiles and W. Leach, *Respiration in Plants*, 3d ed., 1952.

## Plant societies

Assemblages of plants which constitute structural parts of plant communities. They may be components in spatial arrangement such as layers, life-form groups, or seasonally or locally prominent populations of plants. There is no agreement as to the precise usage of the term beyond the generally accepted notion that it should be used for structural vegetation elements of a rank below or within the plant community as a whole. A hierarchy of progressively larger units in vegetation structure with an example is as follows:



Societies can be defined on the basis of structure, dominance, season, and life form.

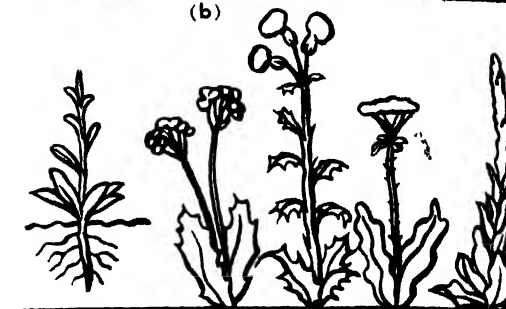
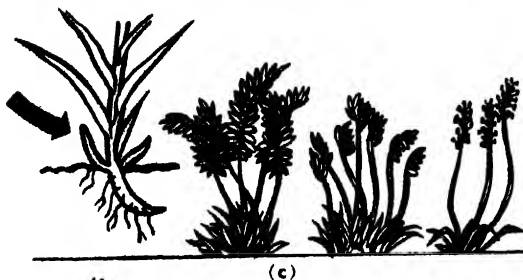
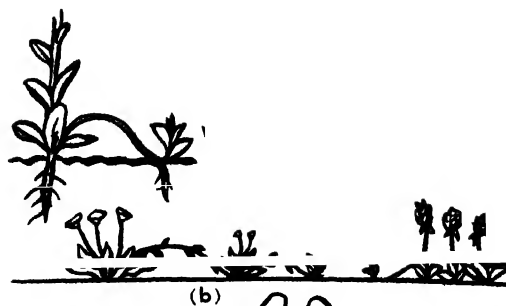
**Structural societies.** These societies are groups of plants within a community, which attain approximately the same height and which bear their foliage at about the same level above ground. They are also known as layer societies or unions. Such societies may form a more or less continuous layer throughout the area occupied by the community. In a forest, for example, there is the canopy society of tall trees, low-tree society, shrub society, herb society, and ground-layer society. These may be refined further if necessary. Some authors have implied or stated that there is a certain cohesion among the component species of a layer society, giving it the status of a community within a community. However, since there is also often a strong interdependence between members of different layer societies, as in the influence of the canopy upon the density of lower vegetation in a forest, such societies should not be regarded as independent entities. A useful set of symbols for the recording of spatial arrangement of vegetation has been proposed by P. Dansereau.

**Dominance societies.** These societies have been defined by J. Weaver and F. Clements as aspect societies or series of stands clearly belonging to a certain community. In addition to the characteristic dominants, these societies possess certain subdominants or codominants of another life form or aspect than the dominant elements of the community. The concept is useful especially in grasslands, marshes, heaths, and other vegetation types in which dominance is an important feature. For instance, prairie communities may be defined on the basis of grass species. Within these communities, certain conspicuous, broadleaved herbs form local aspect societies. The extent and development of such societies have been used as indicators of the recent history of grassland stands, especially with regard to climatic fluctuations.

**Seasonal societies.** Weaver and Clements also applied the term society more precisely to groups of plants which determine the seasonal aspects of plant communities. Examples of such seasonal societies are the carpet of trillium, bloodroot, dog-tooth violet and spring beauty in deciduous forests in the spring, the asters and goldenrods of abandoned pastures in late summer, and the masses of short-lived annuals which suddenly develop after spring rains in the western deserts. Such societies are structural units by virtue of the uniform timing of phenologic response of the plant species involved.

Within a community there may be a progression of seasonal societies as prevernal → vernal → estival → serotinal → autumnal → hiemal. Each society makes its own demand upon the resources of the habitat. Therefore, a full understanding of a community requires observation of all its seasonal aspects.

**Life-form societies.** Plants in a community which have their permanent vegetative axes and



Analysis of a hawthorn-crabapple community in term of life forms (with seasonal correlations). With each society a diagram on the left shows overwintering structure (arrow pointing to bud). (a) Society of low

deciduous trees. (b) Society of sod-forming graminoids (c) Society of rosette-forming hemicryptophytes. (d) Society of stoloniferous chamaephytes. (e) Society of rosette biennials. (f) Structure of the entire community

buds at the same level in or above the soil, constitute life-form societies or synusiae. Members of such societies are therefore subject to similar growth conditions and frequently develop according to a similar pattern. The illustration shows how a plant community may be analyzed structurally using life forms as a criterion to distinguish societies, such as the following: (1) society of clumped, low, deciduous trees (hawthorn); (2) society of sod-forming graminoids flowering in midsummer (bluegrass, redtop, and so forth); (3) society of short-rhizomatous hemicryptophytes with winter rosettes and autumnal flowering period (goldenrod, aster); (4) stoloniferous chamaephytes with winter rosettes, flowering in early summer (pussytoes, cinquefoil, strawberry); and (5) society of biennial rosette plants (wild carrot, thistle, mullein).

Such divisions into societies are useful because they demonstrate the arrangement in space of the aerial and underground parts of the species in the community as well as the timing of their vegetative and reproductive cycles.

From the enumeration of criteria it is evident (1) that all types of societies, regardless of the criteria used to define them, have certain common features; and (2) that there is a lack of agreement regarding the precise meaning and definition of the term "society." Pending a definitive, internationally acceptable vocabulary for ecology, the word society should be used only with a qualifying adjective, as seasonal society or structural society. Where another term with a more specific meaning is available, as union for structural society, or synusia for life-form society, the word should be avoided.

The society concept remains useful to ecologists in the analysis of vegetation, as a general term for

[K.L.E.]

Bibliography: S. A. Cain and G. M. de Oliveira Castro, *Manual of Vegetation Analysis*, 1959; J. R. Carpenter, *An Ecological Glossary*, 1956; J. E. Weaver, *North American Prairie*, 1954; J. E. Weaver and F. E. Clements, *Plant Ecology*, 2d ed., 1928.

## Plant taxonomic literature

Taxonomy is basically descriptive, and its literature is voluminous. It is found in all sorts of publications, from large volumes to small pamphlets, from single works to articles in periodicals. To further complicate the situation, every nation has had its taxonomists and the literature has appeared in many languages. Therefore, great dependence must be placed upon significant indexes to taxonomic literature, floras, and general works of a broad scope.

**Indexes.** One of the most helpful indexes in plant taxonomic literature provides references to scientific names of seed plants. This is the *Index Kewensis plantarum phanerogamarum*, inspired and made financially possible by a gift from Charles Darwin to the Kew Botanical Gardens in England. The original work, published in 1893–1895, consisted of an alphabetical list of genera published from 1753, the date of Carolus Linnaeus' *Species plantarum* (see PLANT NAMES) down to 1885. Under each generic name was given, in alphabetical order, every specific epithet known to have been published in that genus, each entry being followed by the name of the author, the place of publication, and the native country of the plant. Eleven supplements were published up to 1953, and the work has become an indispensable reference for all plant taxonomists.

**Other indexes.** *Genera siphonogamarum*, published in Berlin in 1907, is a list by C. G. Dalla Torre and H. Harms which gives orders, families, and genera of seed plants, arranged systematically; the genera are numbered consecutively from 1, (*Cycas*) to 9629, *Thamnosieris*, and these numbers are used by some curators as a basis for filing herbarium material (see HERBARIUM).

*Thesaurus literaturae botanicae*, second edition prepared by G. A. Pritzel, Leipzig, 1872, gives a list of important titles of botanical works up to the date of its publication.

**General works of a broad scope.** No single work treats of every plant species on earth. Numerous contributions, however, deal with significant segments of the plant kingdom. Some of these are:

*Prodromus systematis naturalis regni vegetabilis*, 17 volumes and 4 index volumes, Paris, 1824–1873. This work attempts to account for all species of dicotyledons. The first seven volumes were written by Augustin Pyramus de Candolle; the remainder were edited by his son, Alphonse, and written by some thirty-five monographers, including Casimir de Candolle, son of Alphonse.

*Die natürlichen Pflanzenfamilien*, A. Engler and K. Prantl, editors, Leipzig, 1887–1915. A 23-volume work including keys, descriptions, and illustrations for families and genera of all plants except bacteria. Summaries are given of information concerning embryology, morphology, anatomy, taxonomy, and paleobotany of each group, with references to selected bibliography.

*Das Pflanzenreich*, Leipzig, 1900—. A work of over 100 volumes, begun by A. Engler, deals with genera and species of vascular plants of the world.

**Floras.** A descriptive manual, or flora, is a systematic treatment of the species of a given area, with identification keys and descriptions of the wild plants of the region (see PLANT KEYS). There are no reference works covering in detail all plants known to exist; even if such reference works were available, they would be too cumbersome for ordinary use. Regional floras, however, make available the details of the plant life of the world, region by region.

In some parts of the world, local vegetation has been studied carefully for many years and few, if any, new species are likely to be found. But in many other regions, manuals are either unavailable or are so antiquated or incomplete as to be of little value. This is especially true of Africa, Asia, and South America, but some parts of Europe, North America, and Australia still lack carefully prepared floras.

Most parts of the United States are covered by regional manuals. For the northeastern states Gray's *Manual of Botany* has been the standard reference work for a century. The 8th edition, edited by M. L. Fernald, appeared in 1950. Another northeastern flora is N. L. Britton and A. Brown's *Illustrated Flora*, a 3-volume work; the third edition, by H. A. Gleason, appeared in 1952. For the southeastern states the principal work is J. K. Small's *Manual of the Southeastern Flora*, published in 1933.

The most extensive flora of the central states is P. A. Rydberg's *Flora of the Prairies and Plains*, published in 1932. For the Rocky Mountain region the most extensive flora is Rydberg's *Flora of the Rocky Mountains and Adjacent Plains*; the 2d edition appeared in 1922. The Pacific states are covered by a comprehensive work, *An Illustrated Flora of the Pacific States*, by L. Abrams and others; the work will include 4 volumes, published in 1923, 1944, and 1951, with the final volume still in preparation.

For southeastern Canada the standard work is Marie-Victorin's *Flore laurentienne*, published in 1935. See PLANT CLASSIFICATION; PLANT KINGDOM. [E.L.C.]

## Plant taxonomy

The systematic classification and arrangement of plants. Plant taxonomy requires a knowledge of categorical concepts such as species, genus, and family, of the rules of both artificial and natural systems of classification, and of international rules of nomenclature controlling specifically the origin and use of a single and universally recognized scientific name. It also includes the making and use of plant manuals, keys, and other classification procedures, together with the employment of approved methods for collecting, systematically recording, and preserving plant specimens. Taxon-



omy involves the application of all other branches of plant science, particularly genetics, evolution, morphology, anatomy, physiology, and ecology.

Modern plant taxonomy is the outgrowth of a long trial and error experience with methods of classification and arrangement. The first scientific botanical writings were produced by the Greek philosopher, Aristotle (384–322 B.C.), and his pupil, Theophrastus (372–287 B.C.). Theophrastus, who is called the Father of Botany, listed the names of nearly 500 plants, and some of these still stand as generic names in modern taxonomy. These early Greek scholars arranged plants in three groups: herbs, shrubs, and trees. Little botanical knowledge was added during the next 1500 years.

In the fifteenth century, men again became curious about plants. During the Middle Ages, plants were classified according to their uses, such as medicinal plants, edible plants, and poisonous plants. About this time there appeared printed botanical books called herbals, which contained the descriptions of plants considered to have medical value. In addition to the descriptions, several herbals were illustrated with woodcuts. Some illustrations were so accurate that today botanists can readily recognize the plants as they were originally pictured. This period, sometimes called the Age of the Herbals, lasted for about two hundred years (1470–1670). During these years botany made the most steady, rapid, and consistent advance recorded up to that time. This period, however, also produced odd and unscientific interpretations, some of which still persist in certain quarters. One of these, the Doctrine of Signatures, maintained that often the medicinal plant was stamped with some clear indication (signature) of its specific remedial power. For example, plants with yellow sap were said to cure jaundice.

As the development of herbals continued, botany advanced from a status of dependence on medicine to that of an independent science. From the classification of plants according to their usefulness to man, there gradually developed an interest in classifying them according to their own natural relationships. Also there arose a growing interest in a more precise system of naming plants so that there might be much less confusion. The publication of *Species plantarum* (1753) by Carolus Linnaeus, a Swedish botanist, marked the end of the old era in plant taxonomy and the beginning of a new epoch. Linnaeus developed a method of classification based primarily on the number of stamens in the flowers. He consistently used the binomial nomenclature introduced by Caspar Bauhin (1560–1624). Linnaeus organized the work of his predecessors and fashioned it into a system by which the average person could identify and name an unfamiliar plant. He recognized natural relationships but believed that only through an arbitrary arrangement could the vast number of known species be presented in a serviceable manner. All these artificial systems of classification were predicated upon obvious superficial characteristics.

However, even before the close of the seventeenth

century, botanists had begun to improve plant taxonomy by making more clear and complete their descriptions of plants, by defining categories more sharply, and especially by attempting to discover a natural basis of classification. The efforts to produce an acceptable natural system were spurred vigorously by the appearance of Charles Darwin's *Origin of Species* (1859) which focused attention upon evolution as a basis for natural relationships.

In 1866, Ernst H. Haeckel coined the term, phylogeny, to designate the genealogical development in phyla of plants and animals. Despite much attention, phylogenetic taxonomy still remains in its early stages because of insufficient information concerning the evolutionary origin and development of the plant kingdom. The true genealogy of plants is not known since this record is buried in the past ages. Hence no scheme of natural classification reveals the complete relationship of all plants. Taxonomists try to place the primitive plants at the bottom and the most advanced plants at the top of the scheme, using morphological similarities and differences as the principal criteria. As investigation brings more and better information, classification methodology should improve, resulting in greater refinement of existing procedures. See PLANT CLASSIFICATION; PLANT KINGDOM.

[P. D. STRAUSBAUGH]

*Bibliography:* Leroy Abrams et al., *An Illustrated Flora of the Pacific States*, 3 vols., 1940–1951; L. H. Bailey, *The Standard Encyclopedia of Horticulture*, rev. ed., 3 vols., 1953; N. I. Britton and A. Brown, *An Illustrated Flora*, 3 vols., 3d ed. prepared by H. A. Gleason, 1952; E. L. Core, *Plant Taxonomy*, 1955; M. L. Fernald, *Gray's Manual of Botany*, 8th ed., 1950; G. H. M. Lawrence, *Taxonomy of Vascular Plants*, 1951; P. A. Rydberg, *Flora of the Prairies and Plains of Central North America*, 1932; P. A. Rydberg, *Flora of the Rocky Mountains and Adjacent Plains*, 2d ed., 1922; J. K. Small, *Manual of the Southeastern Flora*, 1933.

## Plant tissue systems

Most plants are composed of coherent masses of cells called tissues. Large units of tissues having some features in common are called tissue systems. In actual usage, however, the terms tissue and tissue system are not strictly separated. A given tissue or a combination of tissues may be continuous throughout the plant or large parts of it.

Although classification of tissue systems may be based on structure or function, the two aspects usually are combined. Plant tissues are primary or secondary in origin. The primary arise from apical meristems, the perennially embryonic tissues at the tips of roots and shoots (see illustration). The primary tissues include the surface layer, or epidermis; the primary vascular tissues, xylem and phloem, which conduct water and food respectively; and the ground tissues. The latter are parenchyma (chiefly concerned with manufacture and storage of food) and collenchyma and sclerenchyma (the two supporting tissues). In the stem and root, the vascular tissues and some associated

those in the vegetative body of the plant. See separate articles for detailed discussion of the various plant tissue systems. [K. ESAU]

*Bibliography:* See PLANT ANATOMY.

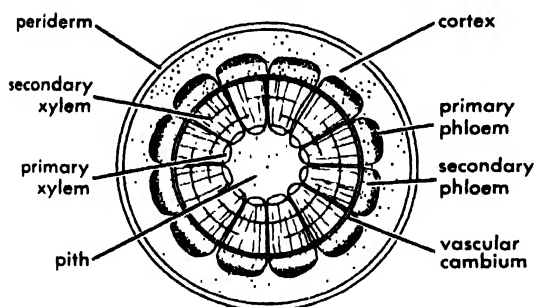
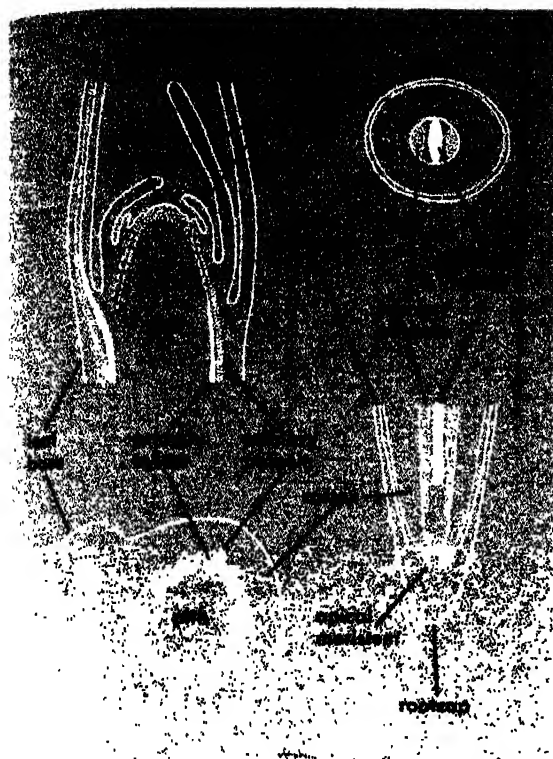
## Plant translocation (organic solutes)

That phase of plant metabolism involving the transport of organic compounds through the vascular tissues (see PHLOEM; XYLEM) of the plant over distances of considerable magnitude relative to the size of the plant, for example, from leaves to roots. See LEAF (BOTANY); ROOT (BOTANY). In arborescent (treelike) species, translocation distances often exceed 100 ft.

Large quantities of organic solutes are translocated in plants. For example, in a single apple tree under favorable conditions, about 150 kg of organic solutes is translocated to the fruit during the course of the growing season. See FRUIT (BOTANY). In addition, large quantities are translocated from the leaves to the growing shoot tips and roots, as well as significant quantities from the roots to the stems and leaves. See MERISTEM, APICAL; STEM (BOTANY). Accordingly, many of the details of the translocation process entail problems of significant concern to plant physiologists and agriculturists from both theoretical and practical viewpoints.

**Importance of conductive tissues.** The relative importance of phloem and xylem tissues as conducting elements in organic translocation varies with the different kinds of organic compounds and, more significantly, with the site of origin or synthesis of the compounds in the plants. Quantitatively, the most important compounds translocated in plants are sucrose and certain closely related oligosaccharides, principally raffinose (see OLIGOSACCHARIDE). Such compounds are translocated predominantly in the phloem, both basipetally and acropetally, that is, in the direction of the root and shoot tips, respectively. During translocation in the stems, a considerable fraction of the sugars in transit may move radially from the phloem into the cortex, xylem, or other stem tissues, and accumulate as starch or other condensed polymers, particularly in the ray cells and in parenchymatous cells immediately adjacent to the tracheae, the main water-conducting vessels (see PLANT TISSUE SYSTEMS; POLYMER).

Despite the proximity of these carbohydrate reserves to the active water-conducting cells in the xylem, only negligible quantities of sugars appear in the sap extracted from the tracheal elements during a major portion of the growing season. Also, numerous ringing experiments, in which the continuity of the phloem tissue is interrupted by the removal of a narrow annular band of tissues external to the xylem, confirm the fact that the xylem tissue plays no significant role in the conduction of sugars. During the dormant season, when the flow of water in the xylem conduits practically ceases, sugars often accumulate in the xylem sap in readily detectable amounts (see PLANT, WATER RELATIONS OF); however, the average maximal concentrations seldom exceed 0.05% for most species,



cross section of dicotyledon stem  
showing primary and secondary tissues

Primary and secondary plant tissue systems.

ground tissue are often treated as a unit, the stele. Ground tissue may be present in the center of the stele (pith) and on its periphery (pericycle). The ground tissue system enclosing the stele on the outside is the cortex. It may have a hypodermis peripherally and an endodermis next to the stele.

The secondary tissues arise from lateral meristems, and their formation is mainly responsible for the growth in thickness of stems and roots. They comprise secondary vascular tissues and the protective tissue called periderm. Secondary growth may build up a massive core of wood, but the outer tissue system, the bark, remains relatively thin because its outer or older part becomes compressed and, in many species, is continuously sloughed off.

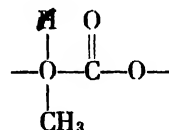
The production of flowers instead of vegetative shoots results from physiological and morphological changes in the apical meristem, which then becomes the flower meristem. The latter, however, produces tissue systems fundamentally similar to

although a concentration as high as 8% has been reported for the sugar maple, *Acer saccharum*. Even with this species, ringing the stem in close proximity to a terminal bud results in such a severe restriction in the supply of organic nutrients to the emergent shoot that growth is greatly inhibited (see PLANT GROWTH). It is evident, therefore, that even at the beginning of the growth period, when the sugar content of the xylem sap is maximal for the season, the importance of the xylem as a translocatory system for sugar is negligible, at least for the species that have been investigated in this respect.

Less data are presently available on the relative importance of xylem and phloem in the transport of nitrogenous compounds, either in inorganic or organic combinations. Chromatographic analyses of the xylem sap of many different species have been reported, and reveal the occurrence of a large number of amino acids, especially aspartic and glutamic acids and the corresponding amides, also occasionally urea and certain related ureides, and alkaloids (see ALKALOID; AMIDE; ACID; ASPARTIC ACID; CHROMATOGRAPHY; GLUTAMIC ACID; UREA). The relative quantities of these compounds vary widely with species. Inorganic forms of nitrogen, especially nitrate-N, occur usually in trace quantities only, or are absent altogether. The concentration of the nitrogenous constituents in the xylem sap, unlike that of soluble carbohydrates, remains moderately high throughout the major portion of the growing season, at an average concentration of about 0.03 *M* glutamine-equivalent in apple, for example. Conversely, the concentration of nitrogenous constituents in the exudate from sieve tubes of the phloem (this exudate is readily collected by making an incision in the bark to the depth of the ac-

tive phloem, as shown in Fig. 1) is usually very low, of the order of 0.001 *M*. These data provide presumptive evidence that the xylem is the predominant channel of transport for nitrogenous compounds, especially from the roots to the leaves, fruits, and stems. On the other hand, data obtained from a number of experiments in which conventional ringing techniques were employed strongly suggest that the phloem is the major transport system. For the present, these and other controversial data bearing on this problem can only be reconciled on the basis that the relative importance of the xylem and phloem in the translocation of nitrogenous compounds varies with species, the ontogenetic (life history) stage of the plant, and various environmental conditions. Redistribution of nitrogenous compounds from the leaves to other parts of the plant is generally considered to occur mainly in the phloem.

The increasing use of systemic spray compounds as insecticides, fungicides, bactericides, and herbicides on plants has centered considerable interest in the study of molecular modifications which increase the absorbability and translocatability of these compounds (see ANTIBACTERIAL AGENTS; FUNGISTAT AND FUNGICIDE; HERBICIDE; INSECTICIDE). Many carbamates, for example, have growth modifying and herbicidal effects, but are not readily translocated when applied to mature leaves (see URETHANE). However, by incorporating the lactic acid group



into the molecule, as in lactic acid *N*-phenylcarbamate or  $\alpha$ -carbodocetoxyethyl-*N*-phenylcarbamate, the translocatability of the molecules is greatly increased. Closely related derivatives without the lactic acid group are not translocatable. Much work remains to be done on the general problem of molecular structure in relation to translocatability. Glucose and fructose, although frequently abundant throughout the plant, are much less translocatable than sucrose (see FRUCTOSE; GLUCOSE; SUCROSE). Chromatographic analyses of the sieve-tube exudate from a considerable number of different tree species have revealed the consistent absence of hexose sugars, except for the occurrence in ash (*Fraxinus*) of mannitol, the alcohol derivative of mannose (see CARBOHYDRATE). Of true sugars, only the nonreducing oligosaccharides of the raffinose family have been found, namely, sucrose, raffinose, stachyose and possibly verbascose. These sugars constitute an ascending series of di-, tri-, tetra-, and pentasaccharides, differing from each other only in the number of included galactose residues (sucrose none, raffinose one, stachyose two, and verbascose three). Sucrose is the dominant constituent in the sieve-tube exudate of all species thus far analyzed, except in ash in which stachyose occurs in the higher concentration. The total stachyose concentration

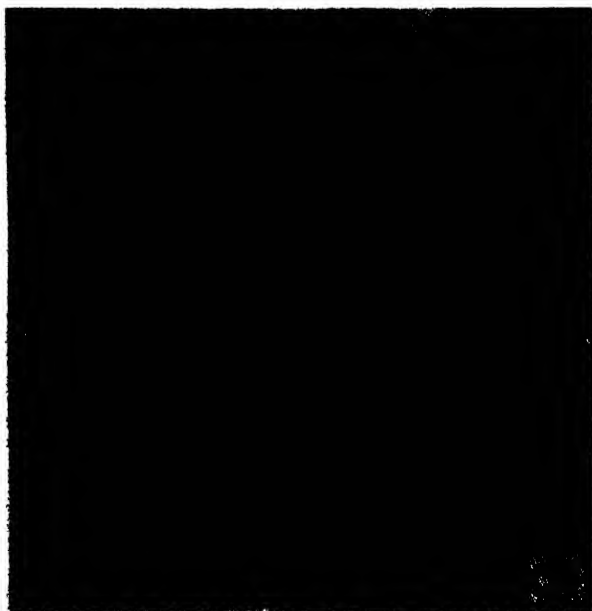


Fig. 1. Droplets of sieve-tube sap exuding from a phloem incision made at the level X-X in the bark of box-elder, *Acer negundo*. (Photograph by K. L. Webb and R. M. Hodgson)



Fig. 2. Stalactites of sugar crystallized from sieve-tube sap exuding from incisions made in the bark of ash trees. (From B. Huber, *Ber. deut. botan. Ges.*, 66(9):340-346, 1953)

from 10 to 25%, varying with species and individual trees and fluctuating diurnally and seasonally. In Sicily, this sugar is harvested from various species of ash trees, especially *Fraxinus ornus*, by tapping the sieve tubes, as shown in Fig. 2. The crystallized exudate forms small stalactites, or sugar-sticks, which may be harvested at convenience.

Whether the sieve-tube sap can be considered with certainty to constitute a valid sample of the organic solutes in transit is not known, but similar conclusions regarding the relative translocatability of sucrose and hexoses have been reached from data based on radiochemical assays (see RADIOCHEMISTRY).

**Requirements for organic nutrients.** The specific requirements of fruit for organic nutrients or nutrient precursors translocated to it from other parts of the plant seem to be relatively few and simple. Excised pollinated ovaries or ovaries of the tomato and gherkin can grow in vitro and form fruits which may occasionally even produce viable seeds on a culture medium containing only sucrose, the usual complement of inorganic ions (nitrates, phosphates, sulfates, calcium, magnesium, and so forth), and water. See FLOWER (BOTANY); PLANT, MINERAL NUTRITION OF; SEED (BOTANY). Although significant differences in the chemical composition of fruit of a given variety may be expected when grafted onto various stock varieties, the actual dif-

ferences in some cases appear to be quite negligible (see GRAFTING OF PLANTS). For example, grafting young sour lemon fruits, without the scion leaves, on to a sweet lemon tree, or vice versa, does not materially change the organic acid composition of the respective fruits when mature (Fig. 3).

This is not to imply, however, that the rootstock has no effect on the chemical composition and quality of the fruit. Consistent differences have been found, for example, in the bitterness of the juice of the Washington Navel orange relatable to the type of rootstock used. Stock-scion relationships are also important in many indirect ways, such as disease resistance and cold hardiness. See PLANT DISEASE CONTROL.

In view of the fact that the preponderance of organic solutes are translocated through the phloem cells, considerable research has been directed toward a clearer elucidation of the physiology of these cells in order that a better understanding may be gained of the mechanism of translocation and of the factors, both environmental and internal, which influence this process. The technological importance of these studies derives from the fact that many crop management practices are, in their final analysis, efforts on the part of the agriculturist to control the rate and direction of translocation of organic compounds from donor to acceptor areas within the plant, as from leaves to roots, or leaves to fruits. See AGRICULTURAL SCIENCE (PLANT).



Fig. 3. Lemon fruit grafted May 4 and photographed May 29. Fruit was 51.4 mm in diameter when grafted and 65.2 mm when mature on December 11, giving a calculated increase in volume of 63%. (From L. C. Erickson, *Science*, 125(3255):994, 1957)

**Mechanism of translocation.** Any mechanism which is proposed to explain the translocation of solutes in the phloem must take into account the following facts:

1. Living cells are essential. Unlike the transport of water and inorganic solutes in the xylem, transport of organic solutes in the phloem can occur only through living cells. Killing the cells in a localized zone of the stem, for example, with steam, boiling water, or hot wax, is as effective a barrier to the transport of phloem-limited solutes as is complete removal of the phloem by ringing. However, the specific cells in the phloem through which the major fraction of the solutes is considered to be translocated, namely, the sieve tubes and other sieve elements, are enucleated cells, and on the basis of various cytohistochemical observations, it is commonly held that the protoplasm of these cells is relatively inactive metabolically compared to that of the other tissues (see CYTOLOGY, PROTOPLASM). Nevertheless, it is apparent that a certain minimum level of metabolic activity is essential (see PLANT METABOLISM).

2. Rate of translocation is rapid. The rate of dry weight increase in fruits, tubers, and fleshy roots reflects the rate of organic translocation, and this quantity varies for different species and environmental conditions from about 0.03 g/hour to in excess of 2 g/hour when averaged over the entire growing season (flowering to harvest). The velocity required to account for these known rates of translocation is about 1–2 cm/min, based on the assumption that the entire cross-sectional area of the sieve tubes (exclusive of the walls) is available for transport. Direct measurements of the velocity of translocation of radioactively labeled compounds through stems provide comparable values. It has been estimated that the rate of translocation is 10,000–100,000 times faster than can be accounted for by diffusion.

3. Translocation is frequently polarized. Growing regions on the plant, especially rapidly developing fruits, exert a strong monopolizing effect on the distribution pattern of organic solutes translocated from the leaves and other areas of synthesis or storage. This polarizing action of fruit is established during the early stages of embryo development, and does not develop in the absence of pollination in nonparthenocarpic fruits. In apple trees, it has been shown that, for a constant ratio of 30 leaves per fruit, the leaves and fruit can be separated by distances up to 10 ft (the maximal distance varying with plant variety) without any loss in the size and quality of the fruit.

None of the current theories which have been proposed to explain the mechanism of solute translocation in the phloem is capable of reconciling all of the known facts regarding this process. The theory that has received the most general acceptance, despite a number of admitted limitations, is that developed principally by E. Münch about 1930, and known as the *Druckstrom*, or pressure-flow mechanism. According to this theory, modified to take

into account newer information and concepts, there is a flow of solution within the sieve tubes in the direction of a positive turgor pressure or hydraulic pressure gradient, that is, from a region of higher to a region of lower turgor pressure within the sieve tube. This pressure gradient is established and maintained osmotically as a result of the metabolic entry and removal of translocatory solutes, principally sucrose, into and from the sieve tubes. The entry of sucrose into the sieve tube at donor sites and its removal at acceptor sites appears to be in many cases against a concentration gradient, and hence must be a metabolically actuated process. It is known that the process of translocation is sensitively influenced by the physiological status of the donor and acceptor organs.

Other theories on phloem translocation, for example, the cytoplasmic streaming hypothesis, attribute a more active role to the individual cells or "elements" of the sieve tube. In brief, however, the present status of the overall problem mainly favors the view that the sieve-tube protoplasm does not possess per se any metabolic machinery capable of driving the translocation process, except for metabolically activated transfer systems which move solutes across limiting membranes at the sites of entry and exit, into and from the sieve tubes. Throughout the major portion of the transport distance, it appears that the solutes are simply swept along at different rates by solvent drag forces resulting from the osmotically actuated flow of water through the sieve tubes. [C A S W]

**Bibliography.** O. F. Curtis, *The Translocation of Solutes in Plants*, 1935; B. S. Meyer and D. B. Anderson, *Plant Physiology*, 2d ed., 1952; E. Münch, *Die Stoffbewegungen in der Pflanze*, 1930; F. C. Steward (ed.), *Plant Physiology: A Treatise* vol. 2, 1959.

## Plant virus

Viruses attack many higher flowering plants, but with the exception of the viruses which attack bacteria or actinomycetes, none has been definitely shown to infect lower forms in the plant kingdom such as ferns, mosses, or algae. Some viruses attack many species of plants in many genera and families; others appear to be limited to a single family.

**Virus symptoms.** Symptoms may occur in any part of the plant, including the roots. The most common symptom is mosaic, that is, patterns of light and dark green areas in leaves (Fig. 1). In most virus diseases, mosaic or other symptoms are preceded by clearing of veins in the leaves of new growth. Some of the older leaves may show primary lesions, for example, yellow spots at points where the virus was introduced. Sometimes there is a mosaic pattern on the stems or fruits as well as on the leaves, or the flower petals may be variegated. The earliest records of plant viruses are Dutch paintings of symptoms in tulip flowers.

**Types of virus diseases.** There are various types of virus diseases. One causes a general lack of





Fig 1 Tobacco mosaic symptoms in a tobacco leaf. The mosaic effect results from the distribution of the light and dark green areas. The white area is due to local, exclusive occupation of the tissues in this spot by a mutant of the virus. (Photograph from L. O. Kunkel)

chlorophyll in new growth, excessive branching, greening of flower petals, and other symptoms (Fig 2). Still other viruses cause galls or tumors, dead streaks, and internal killing of specific tissues (Fig 3). Individual plant cells may be reduced in size killed, or stimulated to enlarge or divide. Cell organelles may show symptoms, and inclusions of various kinds may occur in the cytoplasm or nuclei (see CELL NUCLEUS; CYTOPLASM). Some of these are crystals of virus (Fig. 4). All tissues may be affected, or the virus may have preferential affinity for parenchyma, phloem, or xylem (see PARENCHYMA; PHLOEM; XYLEM).

Plant virus symptoms are markedly influenced by environmental conditions such as temperature, light, and mineral nutrition. When infections are inapparent in the ordinary environment, the plant

involved is referred to as a symptomless carrier; and cultivated plants, weeds, or other wild plants of this sort may be among the important reservoir sources of virus for a particular disease.

**Transmission of viruses.** Fortunately, few viruses are transmitted through seeds or pollen of plants. However, because of the usual systemic nature of the infection, most viruses are perpetuated in plants which are propagated vegetatively by such means as cuttings, bulbs, roots, and grafted shoots (see GRAFTING OF PLANTS; REPRODUCTION, PLANT). A great many cultivated plants, for example, potatoes and sugar cane, are propagated vegetatively, and it is in these that viruses cause the greatest losses.

Almost all plant viruses can be transmitted by grafting a portion of a diseased plant onto a susceptible healthy plant. This is one of the most sensitive methods of ascertaining whether a virus may be involved in a disease. A few viruses are transmitted by any action releasing a minute amount of sap from a diseased cell and introducing it into a healthy cell, for example, wind rubbing leaves together, fingers or tools touching plants. Some are transmitted by mites and some by soil nematodes (see ACARINA; NEMATODA). However, most plant viruses are transmitted by insects, a few by biting insects such as beetles, but most by sucking insects, such as aphids and leafhoppers. Almost all of them are transmitted by only one kind of insect (see ENTOMOLOGY, ECONOMIC; INSECTA).

There are several kinds of relationships between the viruses and their insect carriers. In one relationship, the insect acquires virus on the tip of its minute sucking tubes during a probe of a diseased plant, and transmits it while making similar probes of healthy plants. In another type, the virus is taken into the body of the insect carrier and only after it multiplies there for 1-2 weeks does the insect begin to transmit it. For the rest of its life



Fig. 2. Proliferating shoots, with little or no chlorophyll, arising near the leaf bases of an aster plant infected by aster yellows virus. The flower parts at the tips are transformed to leaflike structures almost devoid of pigment. The older leaves are yellowed; the more mature the leaf when the disease first develops, the less the effect.





Fig. 3 Tumors scattered over the roots and massed at the crown of a sorrel plant infected by wound-tumor virus.

the insect may continue to transmit the virus. Some of these insects are known to be diseased by the virus and pass it to their young through the egg. However, none of the plant viruses is known to attack higher animals.

**Structure and chemical composition.** Viruses have various shapes—rods, flexuous filaments, polyhedra, spheres, or other forms—all of which are too small to be seen with the light microscope but have been resolved by the electron microscope (Fig. 5). The principal components of viruses are protein and nucleic acid, the latter carrying the hereditary determinants. The single strand of nucleic acid occurring in each tobacco mosaic virus particle, if separated from the protein without being broken, is infectious, whereas the protein is not. The nucleic acid strands and subunits of the protein can be recombined in the test tube to reconstitute virus particles. The nucleic acid can also

be chemically changed in the test tube to produce mutants of the virus. These discoveries have contributed to our basic understanding of heredity (see *MUTATION; NUCLEIC ACID; VIRUS*).

Mutations of most plant viruses occur spontaneously and are readily isolated in the laboratory. In nature such mutations also occur but here strains showing greater differences from each other are also found which, for their development, probably require evolutionary intervals not available in the laboratory.

When a virus is introduced into a plant, the protein apparently separates from the nucleic acid. Freshly formed free nucleic acid and protein subunits appear in the cells before the completed virus particles.

The infective agent is assumed to spread through the microscopic protoplasmic strands (plasmodesmata) connecting one cell with another through the cellulose walls (see *CELL STRUCTURES*). Passage through the parenchyma cells is slow but transport through the vascular elements (food- and water-conducting cells) is rapid.

**Economic importance.** Losses from plant viruses vary from the destruction between 1936 and 1946 of 7,000,000 orange trees by tristeza disease in the state of São Paulo, Brazil, to the insidious annual

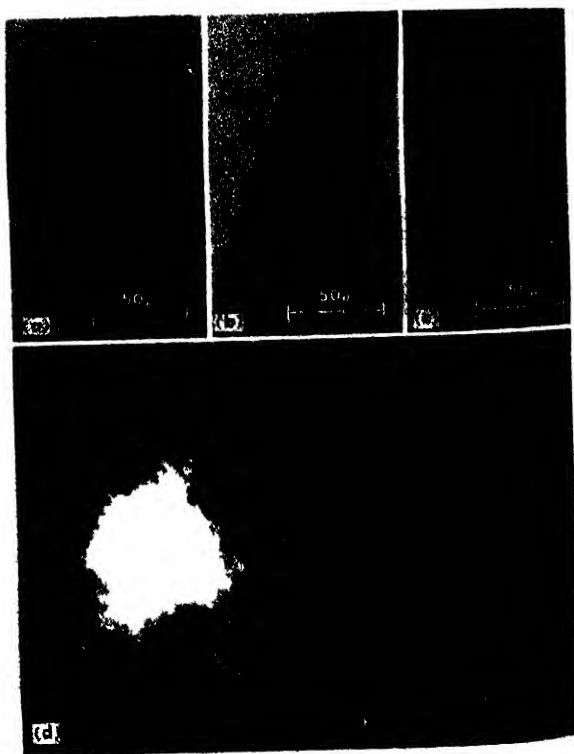


Fig. 4. (a) Crystalline inclusion bodies of tobacco mosaic virus photographed in a living cell of a tobacco plant. (b) The same cell and crystalline inclusion after freeze-drying. (c) The frozen-dried crystal removed from the cell. (d) A partially dissolved fragment of the same inclusion disclosing that it consists of tobacco mosaic virus particles. (Photographs from R. L. Stearn and R. C. Williams)

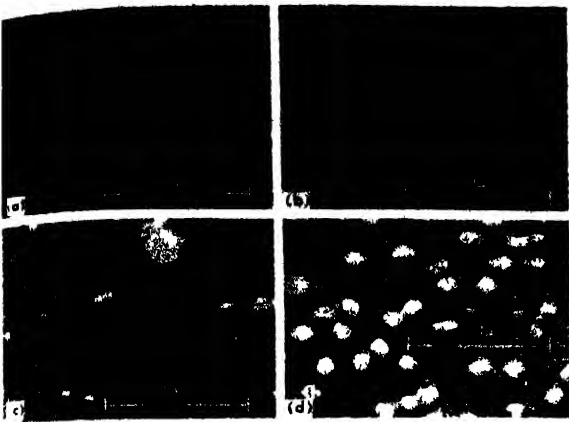


Fig 5 Four forms of plant viruses, approximately the same magnification. (a) Tobacco mosaic virus, a straight rod. The infectious particles are probably close to a standard length (photograph from J. Brandes). (b) Beet yellows virus, a flexible filament, not related to aster yellows virus (photograph from J. Brandes). (c) Wound-tumor virus. Examination suggests that the particles are polyhedral. The large particle is a latex sphere (photograph from M. K. Brakke, A. E. Vatter, and L. M. Black). (d) Under the electron microscope, potato yellow-dwarf virus assumes various shapes approximating balls, sausages, and deflated balloons (electron photograph from M. K. Brakke and A. E. Vatter)

loss caused in certain potato varieties that are infected by one or more viruses producing no noticeable symptoms. Control measures include quarantines, breeding for disease resistance, killing weeds and insect carriers, inspection services, certified seed, and many other practices that have been of vital importance in maintaining crop yields and quality. See PLANT DISEASE; PLANT DISEASE CONTROL.

[L.M.B.]

**Bibliography:** F. C. Bawden, *Plant Viruses and Virus Diseases*, 3d ed., 1950; K. M. Smith and M. A. Lauffer (eds.), *Advances in Virus Research*, vols. 1-8, 1953-1961.

## Plantaginales

A small order of the plant subclass Dicotyledoneae having a single family with 3 genera and 200 species of cosmopolitan range. The plants are herbaceous. The leaves are in rosettes, the inflorescence is scapose, capitate or spicate, and the fruit is a circumscissile capsule, meaning that it opens along a regular, transverse, girdling line. Most of the species are weeds, but the fleawort (*Plantago psyllium*) is cultivated in Spain and France for psyllium seeds used as a laxative. See DICOTYLEDONEAE; EMBRYOPHYTES; PLANT KINGDOM.

[P.D.S.]

## Plasma physics

That field of physics which relates to the study of highly ionized gases. A gas which is composed of a nearly equal number of positive and negative free charges (positive ions and electrons) is called a plasma after the original definition by I. Langmuir.

Because it is composed of charged particles, a plasma exhibits many phenomena not encountered in ordinary gases. In addition to their importance in many new areas of applied science, these effects are evident in astrophysical phenomena (most of the matter in our universe exists in the plasma state), both in stellar atmospheres and in interstellar space. Plasma phenomena have also been observed in the tenuous ionized gases of the earth's outer atmosphere. See COSMIC ELECTRODYNAMICS.

Practical interest in plasma physics arises from various applications of gas discharges, from the study of electron beams in electron tubes, and from the new research fields of ultra-high-temperature processes and controlled fusion (see FUSION, NUCLEAR; PINCH EFFECT). In these latter experiments millions or even hundreds of millions of degrees kinetic temperature are required (kinetic temperature is defined later). To obtain a comprehensive physical picture of plasma it is necessary to consider its behavior from two different aspects:

1. The microscopic picture, relating to its particlelike properties, such as the effects of interparticle collisions in producing diffusion and other transport phenomena, ionization, x-radiation, and other particulate processes. In this picture a plasma exhibits properties some of which are much like those of any gas.

2. The macroscopic picture, where the collective or fluidlike properties are most evident. These properties include conduction of electricity, propagation of various kinds of waves, and ability to support classes of unstable and turbulent behavior peculiar to conducting fluids. Many of the macroscopic behavioral properties of plasma are related to the general field of magnetohydrodynamics (see MAGNETOHYDRODYNAMICS).

The charged particles of a plasma interact with each other through the electrostatic or Coulomb field with which each is surrounded. On the microscopic scale, these electrostatic fields give rise to localized attractive or repulsive forces between the particles as they pass near to each other, resulting in mutual deflection. On the macroscopic scale, the summation of the many infinitesimal electrostatic and magnetic fields produced by the moving plasma particles results in a smeared-out or averaged electromagnetic field. The plasma then reacts collectively, that is, as a conducting fluid, to the total electromagnetic field in which it is immersed. This field consists of the combination of the plasma electromagnetic field and any externally imposed fields. The coupled nature of the plasma motion and the electromagnetic field in which it moves is the source of most of the complexity of plasma behavior.

### CRITERION FOR PLASMA PHENOMENA

A length scale which approximately divides the microscopic domain from the macroscopic domain in a plasma is the so-called Debye screening distance  $\lambda_D$ . It can be shown that as long as the distance between two passing particles is appreciably

Values of  $\lambda_D$  for typical densities and temperatures

$n_e$ , electrons/cm <sup>3</sup>	$\lambda_D$ , cm				$d$ , cm ( $1/n_e$ ) <sup>1/3</sup>
	$T_e$ (°K) { 10 <sup>5</sup> (8.6 ev)	10 <sup>6</sup> (86 ev)	10 <sup>7</sup> (860 ev)	10 <sup>8</sup> (8.6 kev)	
10 <sup>8</sup>	0.22	0.69	2.2	6.9	$2.1 \times 10^{-3}$
10 <sup>10</sup>	0.022	0.069	0.22	0.69	$4.8 \times 10^{-4}$
10 <sup>12</sup>	$2.2 \times 10^{-2}$	$6.9 \times 10^{-3}$	0.022	0.069	$1.0 \times 10^{-4}$
10 <sup>14</sup>	$2.2 \times 10^{-4}$	$6.9 \times 10^{-4}$	$2.2 \times 10^{-3}$	$6.9 \times 10^{-3}$	$2.1 \times 10^{-5}$
10 <sup>16</sup>	$2.2 \times 10^{-6}$	$6.9 \times 10^{-6}$	$2.2 \times 10^{-4}$	$6.9 \times 10^{-4}$	$4.8 \times 10^{-6}$
10 <sup>18</sup>	$2.2 \times 10^{-8}$	$6.9 \times 10^{-8}$	$2.2 \times 10^{-6}$	$6.9 \times 10^{-6}$	$1.0 \times 10^{-6}$

less than  $\lambda_D$ , normal Coulomb attraction or repulsion will exist and one can define the encounter as a simple collision, to which the ordinary laws of particle dynamics apply (see COLLISION). However, if the minimum distance of approach of two particles is greater than  $\lambda_D$ , the collective motions of the surrounding plasma electrons induced by the passage of the particle will be such as to screen the test particle from feeling the influence of the other particle (or any others beyond the distance  $\lambda_D$ ).

The length  $\lambda_D$  depends on the density  $n_e$  and the kinetic temperature  $T_e$  of the plasma electrons. It is usually defined through the relationship

$$\lambda_D = \sqrt{kT_e / 4\pi n_e e^2} \quad \text{cm} \quad (\text{cgs units}) \quad (1)$$

where  $e$  is the charge on the electron and  $k$  is Boltzmann's constant.

Values of  $\lambda_D$  for typical plasma densities and electron kinetic temperatures of interest in laboratory experiments are listed in the table. Kinetic temperature refers to a measure of temperature in terms of the kinetic energy of random motion of the particles of gas. In a maxwellian gas the mean kinetic energy  $\bar{W} = \frac{3}{2} kT$ , where  $T$  is the absolute temperature in degrees Kelvin (see KINETIC THEORY OF MATTER). A convenient measure of  $W$  is the electron volt, so that kinetic temperature is often measured in electron volts: 1 ev kinetic temperature =  $kT = 11,600^\circ\text{K} = \frac{2}{3} \bar{W}$ .

In the last column of the table, the approximate mean particle separation  $d = (1/n_e)^{1/3}$  is given for comparison. Except at the highest densities,  $\lambda_D$  is seen to be substantially larger than  $d$ , corresponding to the fact that many particles are contained within a Debye sphere, so that each particle lies within collision range of many other particles at any given time. This is of importance to the understanding of certain collision effects in a plasma.

The numerical value of  $\lambda_D$  provides an important criterion by which to decide whether in a plasma of given size collective phenomena are to be expected. Certainly, if the over-all dimensions of a region containing plasma are small compared to  $\lambda_D$ , only simple collisional or single-particle behavior is to be expected, the plasma will behave as an ordinary low-density gas, and collective processes will not be important. Conversely, if the

dimensions of a plasma region are very much larger than  $\lambda_D$ , the possibility exists for collective plasma phenomena. Thus, as seen from the table, in the laboratory, where dimensions are measured in centimeters, plasmas with electron densities less than about  $10^{12}$ – $10^{14}$  cm<sup>-3</sup> would not be expected to exhibit collective behavior. These are particle densities which are typically encountered in conventional particle accelerators. On the other hand in the earth's upper atmosphere,  $\lambda_D$  would be much smaller than other typical dimensions for all electron densities higher than a few electrons per cm. In such cases collective effects could be expected to be possible even at the lowest particle densities encountered.

The condition just given admits of another simple and useful physical interpretation. Consider a sphere of plasma with radius  $\lambda_D$ , and assume that nearly all of the electrons of the plasma are removed to infinity from this region. The removal of these electrons will leave an uncanceled positive charge and a resulting radial electric field. From the definition of  $\lambda_D$  it is then easily shown that in this case the energy necessary to remove the additional electrons from the plasma is (within a factor of two) equal to their mean kinetic energy. It follows that if  $\lambda_D \ll$  the plasma dimensions, any influence which tends to separate the bulk of the plasma charges by a distance greater than  $\lambda_D$  will give rise to strong electrostatic restoring forces which will prevent any further separation. Conversely, if  $\lambda_D \gg$  the plasma dimensions, the electrostatic forces arising from even a complete separation of charge will have little influence on the motions of the individual charges so that collective effects will be unimportant.

#### CREATION OF A PLASMA

To create a plasma in the laboratory it is usually necessary to start with an ordinary gas, at a small fraction of atmospheric pressure, and then to heat it by electrical or other means until the mean kinetic energy of the gas particles becomes comparable to the ionization potential of the gas (see IONIZATION POTENTIAL). Mutual collisions of the gas particles will then result in a cascading ionization of the gas. Since ionization potentials are always several volts, such effects are only important at kinetic temperatures of several electron volts, so that the threshold temperature for most plasma

experimentation is 50,000–100,000°K, and ranges up to tens or hundreds of millions of degrees

Returning to the question of producing the plasma by ionization of a gas, one recalls that such processes occur in ordinary gaseous discharges, for example, in fluorescent lamps. However, in such discharges, the ions and electrons of the plasma are continually and rapidly being cooled and recombined by contact with the chamber walls, so that the temperature is low and the state of ionization is only partial and can be maintained only by a large continuous input of energy. A necessary alternative is to find some means of electromagnetic confinement of the plasma, once created, so that its particles cannot touch the chamber walls. Effective confinement is the prime objective of high-tem-

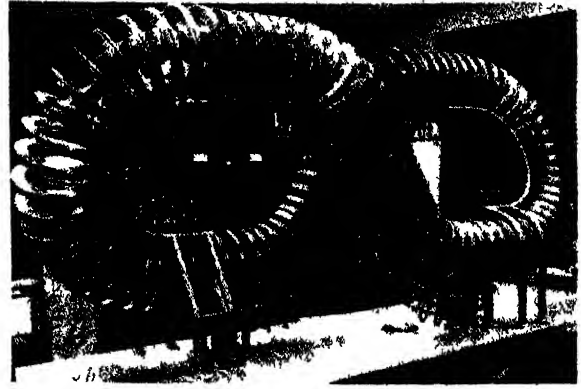


Fig 2 Stellarator. The figure-8 shape prevents certain transverse drifts of the plasma

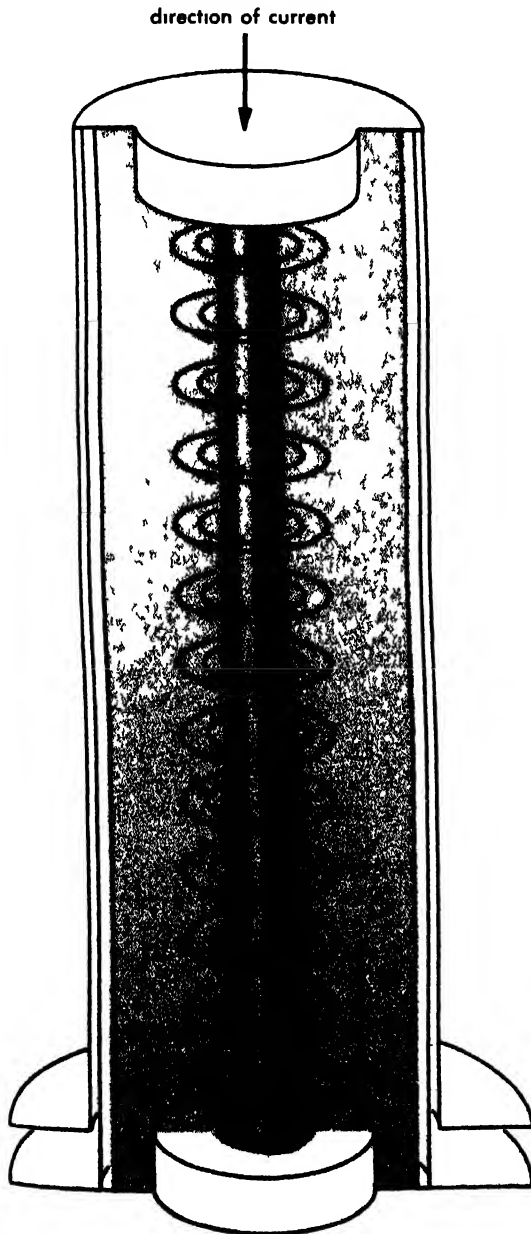


Fig 1 Pinch effect. Closed rings around plasma represent magnetic lines of force.

perature plasma research and is the basis for hopes of achieving controlled nuclear power from nuclear fusion reactions in a hot plasma. Without such means, a high temperature plasma, even if it could be created, would have only a fleeting existence, typically less than a millionth of a second.

**Magnetic confinement.** The pinch effect represents the simplest example of magnetic confinement. If a strong electric current is passed through the body of the plasma, a self-magnetic field configuration of closed rings (as shown in Fig 1) is created. Parallel current elements attract, that is, the body force  $\mathbf{J} \times \mathbf{B}$  ( $\mathbf{J}$  = current density,  $\mathbf{B}$  = magnetic field strength) is always inwardly directed, thus constricting the plasma. From another point of view, the magnetic lines of force, behaving as elastic bands, surround the plasma column and provide a hoop stress which resists the outwardly directed kinetic pressure of the plasma. The current  $I$  required to balance the plasma pressure is given by the Bennett relationship

$$I^2 = 2NkT \quad (2)$$

where  $I$  is in amperes,  $N$  is the total number of particles per lineal centimeter of the pinch,  $k$  is Boltzmann's constant, and  $T$  is the kinetic temperature of the plasma.

Confinement of a plasma by externally generated magnetic fields is also possible. Two generic cases are distinguishable:

- 1 The Stellarator (invented by L. Spitzer, Jr.) exemplifies confinement in an endless tube of modified toroidal form (Fig 2). In the Stellarator a strong magnetic field parallel to the tube walls, generated by external coils, prevents the plasma from touching the chamber walls. Since a longitudinal magnetic field has little influence on plasma motion along the lines of force, it is necessary to close the tube on itself. To avoid certain transverse drifts of the plasma which would be present if the Stellarator were constructed as a simple torus, the field lines are given a helical twist, either by means of special auxiliary coils, or, as in the first Stellarators, by making the tube in the form of a figure 8.

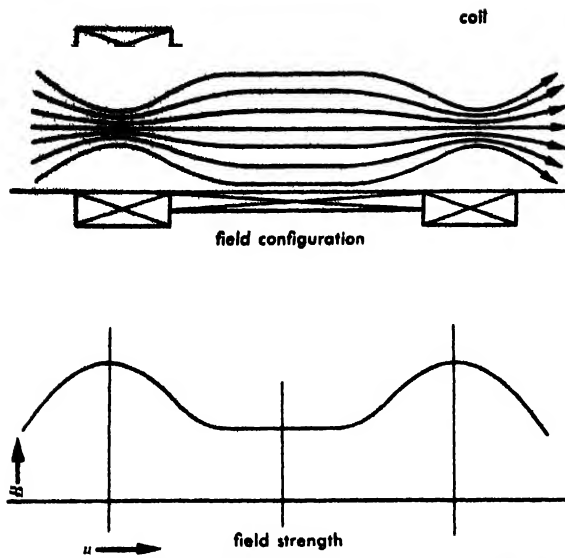


Fig. 3. Magnetic mirror. The helix in the upper part represents the path of a particle in the plasma. The curve below represents the magnetic field strength  $B$  along the path  $u$  in the tube.

2. The "mirror machine" (invented by R. F. Post) allows confinement in a tube with ends "plugged" by magnetic mirrors (Fig. 3). A magnetic mirror is merely a localized region where the magnetic field is made much stronger than average, so that charged particles tend to be reflected as they approach the mirror. The magnetic mirror effect is well known from cosmic-ray physics, where it is encountered in the reflection of cosmic-ray particles by the earth's magnetic field. It has more recently been identified as the explanation of the trapped particle or Van Allen radiation belts of the upper atmosphere, where the earth's dipole magnetic field creates a natural mirror machine. See COSMIC RAYS; VAN ALLEN RADIATION.

**Pressure balance and diamagnetism.** The magnetic confinement of a plasma can be identified with a diamagnetic effect associated with the plasma; that is, the introduction of a plasma into a region containing a magnetic field tends to weaken the magnetic field in that region (see DIAMAGNETISM). This effect is readily seen in the pressure balance relationship for a confined plasma. If a (usually) small effect arising from the curvature of magnetic lines is neglected, this relationship is

$$\nabla \left( p_{\perp} + \frac{B^2}{8\pi} \right) = 0 \quad (3)$$

where  $p$  is the local value of the component of the plasma pressure perpendicular to the lines of force, and  $B$  is the local field strength. This expression, which shows the constancy of the sum of the plasma pressure and the magnetic pressure, can be integrated to yield

$$p_{\perp} + \frac{B^2}{8\pi} = \text{const} = \frac{B_0^2}{8\pi} \quad (4)$$

where  $B_0$  is the strength of the magnetic field just outside the plasma. There exist an infinite number of allowed solutions to this equation, so that other circumstances, such as diffusion, will dictate the actual equilibrium solution in any given physical case.

It is clear from Eq. (4) that the plasma pressure can never exceed  $B_0^2/8\pi$ . This value corresponds to a complete diamagnetic exclusion of the magnetic field from the plasma. Other circumstances, such as plasma instabilities, may impose additional limitations on the plasma pressure, leading to values less than  $B_0^2/8\pi$ . A convenient representation of Eq. (4) can be given in terms of a parameter  $\beta$ , which is defined as the ratio of plasma pressure to externally applied magnetic pressure:

$$\beta = \frac{p_{\perp}}{B_0^2/8\pi}$$

In terms of  $\beta$ , Eq. (4) becomes

$$\beta = \left[ 1 - \left( \frac{B}{B_0} \right)^2 \right] < 1 \quad (5)$$

The parameter  $\beta$  measures the effect which the plasma can have on the applied field. If  $\beta \ll 1$  the applied field will be only slightly affected by the presence of the plasma.

The diamagnetic effect of a plasma arises from persistent electric currents flowing throughout its volume. In fact, the magnetic confining force is just the body force  $\mathbf{F} = \mathbf{J} \times \mathbf{B}$ . But the existence of currents in the plasma must imply a dissipation of energy. Thus, unless maintained, the confining currents will decay with time, so that plasma and magnetic field will gradually intermingle, leading to eventual escape of the plasma. Magnetic confinement of a plasma is therefore necessarily a transient process, with a time scale set by the electrical conductivity of the plasma. This situation is analogous to the slow penetration of a suddenly applied magnetic field into a large metallic conductor. The application of the field induces eddy currents in the surface of the conductor which decay with time, leading to the eventual penetration of the field.

The effectiveness of a magnetic field in confining a hot plasma for a long time depends on the electrical resistivity of the plasma. This is given theoretically by the approximate expression

$$\rho \cong \frac{7.6 \times 10^4 Z}{T_e^{3/2}} \text{ ohm-cm} \quad (6)$$

where  $Z$  is the mean ionic charge. Theoretically  $\rho$  is independent of density, a result which cannot be expected to hold at very low plasma densities. Note also that plasma has a negative temperature coefficient of resistivity. At  $10^7$  °K, a hydrogenic plasma has a theoretical resistivity about as low as pure copper at room temperature, and at temperatures of  $10^8$  °K or higher it is much less.

# PARTICLE DYNAMICS IN PLASMAS

Magnetic confinement, diffusion of a plasma across a magnetic field, and other characteristics of plasma can be better understood by returning to the microscopic picture. Each charged particle of the plasma moves in the smoothed-out electromagnetic field arising from the combination of the external applied fields and the electric and magnetic fields produced by the plasma itself. In many cases collisions are infrequent and the fields are well enough known to allow prediction of the particle motions by relatively simple techniques.

The equation of motion of a charged particle of mass  $M$ , velocity  $\mathbf{v}$ , and charge  $e$  in an arbitrary electric and magnetic field is

$$\mathbf{F} = M \frac{d\mathbf{v}}{dt} = e \left( \mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B} \right) \quad (\text{cgs units}) \quad (7)$$

where  $c$  is the velocity of light.

In many cases of practical interest  $\mathbf{E} \ll \mathbf{B}$ , and  $\mathbf{B}$  will be relatively slowly varying in time and space. In such cases the solutions to Eq. (7) represent orbits which are approximate helices (coil-spring-shaped), corresponding to a rotation of the particle around a line of force superposed on a translational motion along the lines of force (parallel or antiparallel to the direction of  $\mathbf{B}$ ). For the component of motion perpendicular to the direction of  $\mathbf{B}$ , Eq. (7) reduces simply to a centrifugal force equation

$$F = \frac{Mv^2}{r_c} = \frac{evB}{c} \quad (\text{cgs units}) \quad (8)$$

that is, a rotation at the cyclotron angular frequency

$$\omega_c = \frac{v}{r_c} = \frac{eB}{Mc} \quad (9)$$

This frequency is clearly much higher for electrons than for positive ions.

Corresponding to these frequencies of rotation, the radius of curvature  $r_c$  of the orbit is also given by solving Eq. (8) for  $r_c$  as

$$r_c = \frac{Mc}{eB} \frac{2W_{\perp}}{M} = \frac{1}{\omega_c} \frac{2W_{\perp}}{M}$$

where

$$W_{\perp} = \frac{1}{2} Mv_{\perp}^2 \quad (10)$$

For the same  $W_{\perp}$ ,  $r_c$  is much smaller for electrons of mass  $m$  than for ions, corresponding to a value of

$$r_{ce} = 3.4 W_{\perp}^{1/2} / B \quad \text{cm} \\ (W \ll mc^2 = 511 \text{ kev, } W \text{ in ev})$$

For ions

$$r_{ci} = 145 (A^{1/2} / Z) (W_{\perp}^{1/2} / B) \quad \text{cm}$$

where  $A$  is the mass of the ion in atomic mass units.

To take an example, if  $A = Z = 1$  (protons),  $W_{\perp} = 10^4$  ev, and  $B = 10^4$  gauss, then  $r_{ci} = 1.45$

cm. At the same energy and field strength,  $r_{ce}$  is only 0.34 mm. In this case  $\omega_{ci} = 9.5 \times 10^7$  rad/sec and  $\omega_{ce} = 1.75 \times 10^{11}$  rad/sec.

Returning to the equation of motion, Eq. (7), one may now precisely define the conditions under which simple solutions are obtained. It is only required that  $\mathbf{B}$  should vary slowly in time compared with  $(\omega_{ci})^{-1}$  and by a small amount, percentage-wise, over an orbit radius  $r_c$ , that is

$$\tau = \left( \frac{1}{B} \frac{\partial B}{\partial t} \right)^{-1} \gg 1/\omega_c \quad \text{sec} \quad (11)$$

$$\lambda = \left( \frac{1}{B} \frac{\partial B}{\partial r} \right)^{-1} \gg r_c \quad \text{cm} \quad (12)$$

These are the so-called conditions of adiabaticity for the particle motion.

If these conditions are satisfied, some very important consequences follow. These are (i) that the motion of any charged particle can be well represented by following the motion of its instantaneous center of rotation or guiding center, (ii) that the guiding centers will move about with respect to the magnetic lines with slow drift velocities predictable from simple laws, and (iii) that many of the salient features of the motion can be prescribed in terms of nearly constant quantities known as adiabatic invariants.

**Drift velocities.** Assumptions (11) and (12) can be used to calculate drift velocities. The simplest and most important of these velocities follows by inspection of Eq. (7). If the plasma particles move in an electric field with a component transverse to  $\mathbf{B}$ , then the guiding center of each particle will drift in a direction perpendicular both to  $\mathbf{E}$  and to  $\mathbf{B}$ , with a velocity  $\mathbf{v}_0$  which is independent of the charge, mass, or energy of the particle. This velocity is given by

$$\mathbf{v}_0 = c \frac{\mathbf{E} \times \mathbf{B}}{B^2} \quad \text{cm/sec} \quad (13)$$

The derivation of Eq. (13) from Eq. (7) follows immediately if it is noted that in a frame of reference moving at velocity  $\mathbf{v}_0$ , the motional electric field just cancels the applied electric field and the particle motions are again helices. It is concluded that an electric field component perpendicular to  $\mathbf{B}$  causes the plasma to move and distort locally so as always to make the electric field vanish in the plasma's own frame of reference. This is just the behavior which would be expected for a compressible, highly conducting gas. Further analysis shows that this property is also identifiable with a strong tendency for the plasma to preserve constant magnetic flux through each of its volume elements as it moves from place to place or is subjected to slowly varying magnetic fields.

The drift velocity  $\mathbf{v}_0$  is also related to the paradoxical situation whereby a jet of plasma can cross through an evacuated region containing a strong transverse magnetic field. If a plasma jet is impelled at velocity  $\mathbf{v}$  into such a region, a very



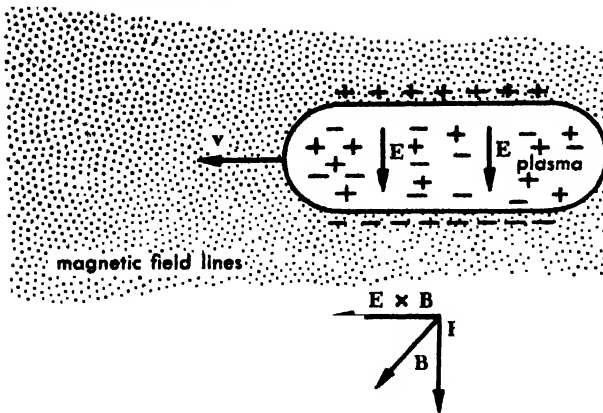


Fig. 4. Schematic illustration of drift of a plasma across a magnetic field in vacuum.

slight separation of charges (Fig. 4) is all that is required to generate an internal polarization electric field in the plasma which maintains the velocity. However, if the plasma passes into a strongly conducting region (another plasma) the polarization fields would not persist, and the motion would stop, the momentum being taken up by the magnetic field or converted into rotation.

In another case, if a plasma is immersed in a magnetic field which is then increased slowly, induction electric fields will appear which will cause the plasma to be compressed toward the magnetic symmetry axis of the system (where  $E = 0$ ). In a uniform magnetic field of circular cross section this would result in a uniform radial compression of the plasma. Throughout the process, the flux through the plasma would remain constant (ignoring diffusion effects).

These illustrations point up the importance of the  $E \times B$  drift velocity. Any situation in which electric fields appear in the plasma because of induction effects or separation will produce this drift. In some cases drifts of this kind can be self-perpetuating (charge separation leading to a drift, leading to more charge separation, and so on), so that a plasma instability results. In other cases the  $E \times B$  drift is accompanied by plasma heating, as in the example of the magnetic compression given earlier.

If the magnetic field in which the plasma particles move is inhomogeneous, other drift motions occur. If there is a gradient of the magnetic field strength perpendicular to the direction of the field, the radius of curvature of each particle will clearly be smaller on the high field side of its orbit than on the low field side, and a slow transverse drift perpendicular to the directions of both  $\nabla B$  and  $B$  will occur. Since oppositely charged particles spiral in opposite directions, this drift will be oppositely directed for electrons and positive ions, so that it produces a tendency for charge separation to occur. The magnitude of the drift is given by

$$v_b = \frac{r_c v_{\perp}}{2} \left( \frac{\nabla_{\perp} B}{B} \right) \quad (14)$$

As the charged particles move in helical paths along the magnetic lines of force they may encounter regions where the flux tubes (bundles of magnetic lines) are curved. While the particles are being guided around these curves by the magnetic field, centrifugal forces will arise which will produce a drift. The centrifugal drift is also oppositely directed for ions and electrons and has the magnitude

$$v_c = v_{\parallel}^2 / R \omega_c \quad (15)$$

where  $v_{\parallel}$  is the velocity of motion of the particle along the lines of force and  $R$  is the local radius of curvature of the magnetic lines.

The centrifugal drift  $v_c$  is an example of a more general gyroscopic kind of drift that can be expected to arise in situations where the plasma particles are subjected to a force which is perpendicular to the local field direction. Another example is the drift velocity  $v_g$  which will occur in the presence of a gravitational field. The magnitude of this drift velocity is

$$v_g = g_{\perp} / \omega_c \quad (16)$$

where  $g_{\perp}$  is the component of gravity perpendicular to the magnetic field. In strong magnetic fields  $\omega_c$  is very small, but it may play a role in geophysical phenomena in the upper atmosphere.

It is essentially impossible to study a confined plasma in the laboratory without encountering magnetic field gradients and therefore stimulating the drifts  $v_b$  or  $v_c$ . Since either of these drifts can give rise to charge separation effects, unless care is taken in choice of the field configuration electrostatic fields within the plasma can be set up which will cause the  $E \times B$  drift  $v_e$  to occur, leading to a rapid escape of the plasma. A classic example is the simple torus with magnetic field lines parallel to the torus walls, as shown in Fig. 5. This geometry, the forerunner of the Stellarator, is characterized by a magnetic gradient which is everywhere inwardly directed (the field at the inner wall of a simple torus is always stronger than at the outer wall, since the line integral  $\oint B \cdot d\mathbf{l}$  is the same for all paths taken around the inside of the torus). Thus, if a plasma is confined in a simple torus, upward and downward drift motions will tend to occur, as predicted by producing a vertically directed electric field and a subsequent common outward drift of the plasma toward the outer wall, that is.

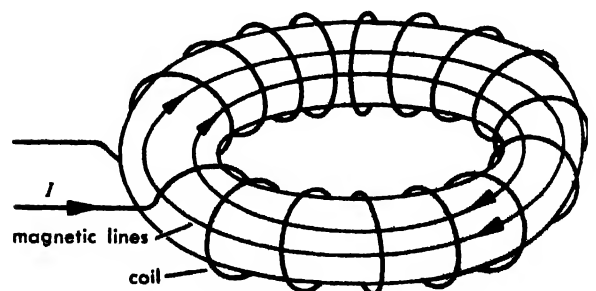


Fig. 5. Magnetic field-lines in simple torus

toward the region of weakest magnetic field. In the first Stellarators, this drift was canceled, on the average, by twisting the torus into a figure 8. Thus, like the famous Möbius strip of mathematics, the inside wall on one curve becomes the outside wall at the opposite curve, so that no net drift results.

**Adiabatic invariants.** Under the adiabatic conditions defined by Eqs. (11) and (12) there are some quantities which are approximate constants of the motion, the so-called adiabatic invariants alluded to earlier. These invariants are useful in predicting many of the details of plasma behavior and have been used as a starting point for erudite theoretical analyses. The simplest of these invariants is  $\mu$ , the magnetic moment associated with the rotation of any charged particle in a magnetic field (see MAGNETON). A moving charge represents a current, and a circling charge therefore represents a circular current element, with which is associated a dipolelike magnetic field similar to that of a simple loop current. By analogy to Lenz's law, this current loop is always diamagnetic; that is, it tends to depress infinitesimally the field strength inside the loop, so that an assembly of charged particles, a plasma, usually exhibits a bulk diamagnetic effect arising from the vector summation of the individual particle diamagnetic effects.

The magnitude of  $\mu$  is given by the expression

$$\mu = W_{\perp} / B \quad \text{ergs/gauss} \quad (17)$$

The magnetic moment  $\mu$  may be expected to be a constant provided the adiabatic conditions (11) and (12) are fulfilled. More specifically, examination of some theoretical work suggests that, except for the effect of collisions, the maximum fluctuations in  $\mu$  occurring as a particle moves back and forth in a varying magnetic field should be roughly representable by the expression

$$\left| \frac{\delta \mu}{\mu} \right| = ac^{b/\epsilon} \quad (18)$$

where  $a$  and  $b$  are constants of order 1, and  $\epsilon$  is either  $2r_c/\lambda$  for spatially varying fields or  $1/\omega_c\tau$  for time-varying fields. It is clear that if  $\epsilon$  is 0.1 or less, the fluctuations in  $\mu$  should be very small. One concludes that there are many plasma situations where the constancy of  $\mu$  between particle collisions should be a valid assumption.

The value of  $\mu$  in plasma calculations can be illustrated by using it to calculate the condition for reflection of a charged particle by a magnetic mirror. Suppose that a spiraling particle approaches a magnetic mirror from a region where the field is  $B_0$ , and is to be reflected when it reaches the point  $M$  where the field intensity is  $B_M$ . From (17), it is seen that

$$\mu = \frac{W_{\perp}(O)}{B_0} = \frac{W_{\perp}(M)}{B_M} = \frac{W}{B_M} \quad (19)$$

The last part of the equation follows since the entire energy resides in rotational motion at the point of reflection. It follows that the condition for re-

flection at a point where the field is  $B_M$  is

$$\frac{W_{\perp}(O)}{W} = \frac{B_0}{B_M} = \frac{1}{R} \quad (20)$$

where  $R$  is called the mirror ratio. Thus, since

$$\frac{W_{\perp}(O)}{W} = \frac{v_{\perp}^2}{v^2} = \sin^2 \theta \quad (21)$$

$\theta$  being the pitch angle of the helix at  $B = B_0$ , the condition that a particle be reflected at or before it penetrates to  $M$  is just that the pitch angle should be greater than a critical angle  $\theta_c$ , where

$$\sin \theta_c = \frac{1}{R^{1/2}} \quad (22)$$

This condition is independent of charge, mass, or total energy of the reflected particle, except as limited by the requirements of the adiabatic assumptions. It is equally easy to show that in the general case, the pitch angle  $\theta$  transforms in accordance with the relationship:

$$\sin \theta(u) = [R(u)]^{1/2} \sin \theta(O) \quad (23)$$

where  $R(u) = B(u)/B(O)$ . Equation (22) follows by setting  $\theta(u) = \pi/2$  (reflection). Equation (23) resembles Snell's law of optics, with  $R^{1/2}$  playing the role of the index of refraction.

Returning to the question of magnetic mirror reflection, it is apparent that if two magnetic mirrors are used, one at each end of a weaker central field, the charged particles of a plasma can be trapped in the "magnetic bottle" between the two mirrors, provided the particles satisfy the pitch angle requirements. This is the mirror machine. If a plasma containing particles with randomly oriented pitch angles were to be suddenly thrust into such a magnetic bottle, all particles with pitch angles less than  $\theta_c$  would immediately escape. The remainder of the particles would be trapped, however, independent of their charge, mass, or energy, and could escape only as mutual collisions deflected them into unfavorable pitch angles. At high temperatures this process is predicted to be quite slow.

Another consequence of the constancy of  $\mu$  relates to plasma heating by magnetic compression. It has already been pointed out that an increasing magnetic field will compress a plasma, which tends to maintain constant flux through its volume. This property is also true for the flux threading each orbit. Since  $\mu = W_{\perp}/B = \text{constant}$ , and  $r_c \sim (1/B) W_{\perp}^{1/2}$ , this implies that  $Br_c^2 \sim W_{\perp}/B = \text{constant}$ , that is, the flux through the orbit circle area  $\pi r_c^2 B$  is a constant. The heating effect follows simply by noting that since  $\mu$  is constant,  $W_{\perp}$  must increase in direct proportion to  $B$ . This is the process of adiabatic compression heating often alluded to in the literature.

To complete the list of adiabatic invariants of plasma motion, two more will be mentioned. Suppose that the plasma particles are trapped in a magnetic bottle and move in a periodic way back

and forth between limits (as in the mirror machine). It can then be shown that the so-called action integral is an adiabatic invariant. The integral is just the line integral of the momentum component parallel to the direction of the field, taken along the path of the guiding center, that is,

$$\oint p_{\parallel} du = A = \text{const} \quad (24)$$

The action integral has the dimensions of an area in  $p_{\parallel}, u$  (phase) space. Its constancy implies that this area is a constant. Physically, it simply means that as a trapped particle moves about in the confinement volume, its guiding center motion must always be such as to keep the phase space area constant (Liouville's theorem; see STATISTICAL MECHANICS). If axial compression occurs, this implies that the axial momentum must increase correspondingly (see Fig. 6). One of the important results of adiabatic theory is to show that the constancy of  $A$  implies that, in the absence of collisions, and for slowly varying fields, the drift motion of trapped particles in a magnetic bottle (such as the earth's magnetic field) is such as to generate an imaginary fixed closed surface for each trapped particle, to which it would be forever bound. The now-famous Argus experiment provided a substantial degree of confirmation of these ideas.

The last adiabatic invariant to be mentioned has already been hinted at. It has been demonstrated that the magnetic flux enclosed by the aforementioned surfaces which are generated as a consequence of the constancy of  $A$  is also an adiabatic invariant. This fact is very useful in analyzing novel or complicated magnetic confinement geometries.

In summary of the particle dynamics of plasmas, it has been shown that plasma particles in a strong magnetic field are generally constrained to move in helical orbits. In the presence of field gradients and other perturbations they tend to drift from line of force to line of force, tracing out closed surfaces as they move about. Throughout

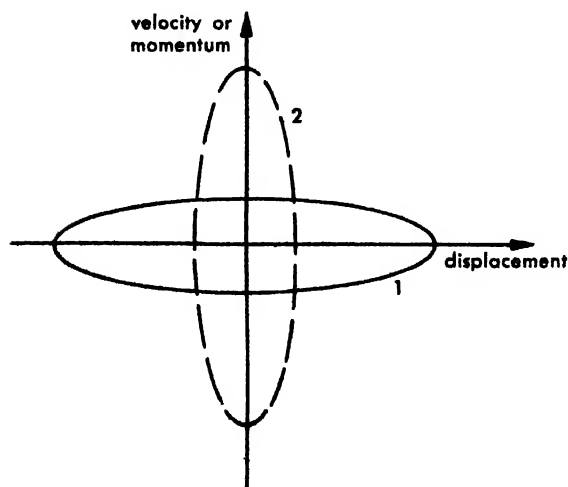


Fig. 6. Schematic illustration of conservation of action integral, Eq. (24).

the motion of the plasma particles, the collective motion will be constrained to be such as to preserve near equality of total positive and negative charge, since any departure from neutrality gives rise to strong electric restoring forces.

**Confinement criteria.** With these facts in mind, the problem of magnetic confinement may now be stated in detail:

1. To immerse an electrically neutral plasma in a magnetic field sufficiently strong that all particle orbits have diameters small compared with the dimensions of the confinement chamber. This ensures that the particles cannot at once touch the chamber walls.

2. To find a configuration of magnetic fields which prevents the rapid escape of the plasma along the magnetic lines and yet does not introduce uncanceled drift motions which would cause it to strike the chamber wall. This requires either that the lines of force shall close on themselves (as in the pinch effect or the Stellarator) or that something akin to the magnetic mirror effect be used.

3. To ensure that the plasma can be sustained in a state of stable pressure equilibrium with the magnetic field, either by use of very strong fields, special field configurations, or special plasma conditions.

The last-mentioned problem is potentially the most difficult one and will be referred to again in the discussion of plasma instabilities.

#### WAVES AND INSTABILITIES IN PLASMAS

Because of the coupled nature of the motion of plasma and its electromagnetic field environment it can support unusual oscillatory or wave motions both stable and unstable.

**Stable wave motions.** A rough subdivision of the stable wave motions is possible in terms of four characteristic "frequencies" or periods of a plasma assumed immersed in a magnetic field. In many typical cases these characteristic frequencies occur in the following order: (i) Interparticle collision frequencies (lowest); (ii) Ion cyclotron frequency; (iii) Electron cyclotron frequency; (iv) The "plasma frequency,"  $\omega_p = \sqrt{4\pi ne^2/m}$  (highest). Only a few of the features of wave propagation will be sketched, the details of plasma wave propagation being remarkably complicated. The reader is referred to the now-voluminous literature on the subject for further details.

At frequencies below (i) a plasma would behave as an ordinary gas and propagate a simple elastic wave. Such waves would be important only for high density, low temperature, or large plasmas.

At frequencies between (i) and (ii) a characteristic plasma wave, the hydromagnetic or Alfvén wave, is possible. This wave resembles that which would be propagated by a loaded elastic string. Here the "strings" are the magnetic lines of force, loaded by the plasma mass (that is, primarily the ions). Satisfying the adiabatic criteria, particles and field distort together (field lines "stick" to the plasma and vice versa) providing a loaded elastic

um for the waves. The Alfvén wave velocity is given by the expression

$$v_A = \sqrt{B^2/4\pi\rho} \text{ cm/sec} \quad (25)$$

where  $\rho$  is the plasma density.

The expression is valid only for frequencies  $\ll (b)$  and for values of  $B$  and  $\rho$  such that  $v_A \ll c$ ; that is,  $B^2/8\pi \ll \frac{1}{2}\rho c^2$  (magnetic energy density small compared with one-half the mass energy density).

At frequencies in the vicinity of (ii), resonance of the wave motion with the ion cyclotron frequency occurs and the waves become highly dispersive.

At frequencies between (ii) and (iv), wave propagation is allowed only for special directions of propagation with respect to the magnetic field, and types of waves are propagated which resemble the ordinary and extraordinary light wave in a birefringent crystal. See CRYSTAL OPTICS.

Above frequency (iv), the plasma propagates simple electromagnetic waves with a phase velocity greater than  $c$ ; that is, it behaves like a medium with an index of refraction less than unity. The index of refraction  $v$  is given by

$$v = \sqrt{1 - (\omega_p/\omega)^2} \quad \omega > \omega_p \quad (26)$$

This property of a plasma can be used to determine its electron density. By measuring the phase shift in the transmission of a microwave beam propagated through the plasma the mean index of refraction can be determined, and from this the density, through use of Eq. (26). The numerical value of the plasma frequency is given by

$$f_p = \frac{\omega_p}{2\pi} = 9 \times 10^3 (n_e)^{1/2} \quad (27)$$

For a density of  $10^{14}$  electrons/cm<sup>3</sup>,  $f_p = 9 \times 10^{10}$  cps, corresponding to a free-space wavelength of 3.3 mm.

The plasma frequency itself represents a classic example of cooperative effects in a plasma. It represents a natural frequency of oscillation of the electrons of the plasma, occurring even in the absence of any magnetic field. It arises from the fact that any net displacement of the plasma electrons with respect to the ions produces a polarization electric field which acts as a restoring force. Thus, if such a displacement occurs, volume oscillations of the electrons about the position of charge neutrality will result. The ions, being much more massive, will remain essentially stationary, like plums in a pudding. If a more detailed analysis is performed, it can be shown that the effect of the thermal velocity of the electrons is to modify slightly the allowed frequency of these electrostatic plasma oscillations, so that a *dispersion* relation results; that is, a frequency-wavelength relationship, showing that such oscillations can occur over about an octave, bounded on the low frequency side by  $\omega_p$ .

**Instabilities.** The instabilities exhibited in a plasma are perhaps its most recondite property. An isotropic isothermal plasma with a maxwellian

particle distribution could not exhibit any instability. However, all plasmas created in the laboratory differ in some respect from such an ideal plasma, possessing a higher degree of order and thus a lower state of entropy. This difference in entropy makes it thermodynamically possible for instability or turbulence to occur, which increases the entropy by introducing a greater degree of disorder. In a magnetically confined plasma, instabilities usually result in the rapid escape of the plasma. A simple rule for the theoretical growth time of many plasma instabilities is simply to divide the over-all dimensions of the plasma by the thermal velocity of the ions. This is usually a very short time for a hot plasma, being of the order of a few microseconds. Understanding and controlling plasma instabilities is clearly a prime target of high-temperature plasma research.

One now distinguishes two general classes of plasma instabilities: (i) configuration space instabilities, in which the instability arises because of order in configuration space (pressure gradients, magnetic field curvature, and so forth); and (ii) velocity-space instabilities, arising from ordering in velocity space (streaming motion, ordered departures from a maxwellian velocity distribution, and so forth). The hydromagnetic instabilities represent the main configuration space instabilities. Such instabilities may occur if there exists an energetically favored interchange or flow of the plasma and the confining magnetic field which will allow the plasma to expand.

When streaming motion of any substantial fraction of the particles of a plasma occurs, as when a heavy electric current is being carried by the electrons (the pinch effect), under some conditions it is possible for an exponentially growing disturbance to occur, drawing its energy from the ordered motion of the stream. The familiar instability of a fire hose has a close analogy in a plasma, to cite one of several examples.

Another type of velocity-space instability, which would feed on a resonant interaction between electron plasma oscillations in a magnetic field and the ordered energy of an anisotropic electron velocity distribution, has been theoretically predicted. A sufficient condition for the plasma to be stable against such unstable oscillations is that

$$(\omega_p/\omega_{ce})^2 < 1 \quad (28)$$

that is, the plasma frequency should be less than the electron cyclotron frequency. The condition can also be written, from the definition of the quantities, as equivalent to the condition

$$(n_e mc^2)/2(B^2/8\pi) < 1 \quad (29)$$

that is, the mass energy density of the electrons must be less than twice the energy density of the magnetic field. Again, it can be rearranged so as to define a critical  $\beta$  value for the electron pressure, since

$$\beta_e = p_e/(B^2/8\pi) = n_e kT_e/(B^2/8\pi)$$

Thus

$$\beta_e < 2(kT_e/mc^2) < 2(T_e/511 \text{ kev}) \quad (31)$$

if  $T_e$  is measured in kev. This is a fairly stringent condition, especially at low electron energies. If one still further manipulates the condition, it can be shown to be equivalent to the requirement that

$$(r_{ce}/\lambda_D) < 1 \quad (32)$$

that is, the mean gyromagnetic radius  $r_{ce}$  of the electrons should be less than the Debye length. It has already been pointed out that cooperative effects cannot occur within distances shorter than a Debye length, so this condition is by no means surprising.

In concluding the discussion of instabilities, it is important to recognize that at the time of this writing only the grossest features of the theory of plasma instabilities have been corroborated experimentally. Since the question of plasma instability is the greatest potential barrier to effective magnetic confinement, the study of instabilities must proceed much farther before even the feasibility of long-time confinement is demonstrated.

#### COLLISION PROCESSES IN PLASMAS

In the absence of instability mechanisms which could destroy a state of order in a plasma, the drive toward a state of higher disorder comes about through interparticle collisions. The dominant interparticle collision process in a high temperature plasma is Rutherford scattering, elastic scattering arising from the mutual Coulomb electrostatic fields of the charged particles of the plasma. Such processes lead to the deflection of and energy exchange between the plasma particles. These collision processes are important, since they determine the basic rate of all collisional transport processes in the plasma. The cross section or effective collisional area  $\sigma$ , for "close" collisions between plasma particles of equal mass and charge can be simply estimated from the classical minimum distance of approach of two equal charges (where mutual potential energy becomes just equal to initial kinetic energy, that is,  $Z^2e^2/r_{\min} = W$ ):

$$\sigma_c = \pi Z^4 e^4 / W^2 \quad \text{cm}^2 \quad (33)$$

However, if the classical minimum distance of approach is compared with the value of the Debye length  $\lambda_D$  given in the table, it will be readily seen that at the usual particle energies  $\lambda_D$  is typically some  $10^6$  times larger. Thus the probability of a particle "colliding" with some particle that is distant yet still within the range  $\lambda_D$  is some  $10^{18}$  times larger than that it should make a simple "hard" close collision. Thus, even though the effect of each individual distant encounter (within  $\lambda_D$ ) is infinitesimal, the statistical sum of their effects is not, and in fact is about 10 times more important than that of close collisions. In the calculation of this effect, the logarithm of the ratio  $(\lambda_D/r_{\min}) = \Lambda$  appears;  $\ln \Lambda \cong 20$  in most cases. One finds for an approximate value of the scatter-

ing cross section owing to distant collisions the value

$$\sigma_d \cong (\pi Z^4 e^4 / W^2) (\ln \Lambda / 2) \quad (34)$$

Written in terms of the particle energy in ev, this becomes

$$\sigma_d \cong 10\sigma_c \cong 6 \times 10^{-13} (Z^4 / W^2) \quad \text{cm}^2 \quad (35)$$

where  $W$  is in ev.

Some quantitative values are of interest. If  $Z = 1$  (hydrogen ions, or singly charged heavier ions), and if  $W = 1$  ev, then  $\sigma_d \cong 6 \times 10^{-13} \text{ cm}^2$ , or some  $10^3$  to  $10^4$  times normal atomic or gas kinetic cross sections. However, if  $W = 1$  kev,  $\sigma_d$  is  $10^6$  times smaller (almost  $6 \times 10^{-19} \text{ cm}^2$ ) and is already much much smaller than atomic cross sections. It can be seen that even disregarding the fourth power dependence on  $Z$ , the range of collision cross sections encountered in hot plasmas may be enormous. It follows that the collision mean free path  $d$  can vary over an even more extreme range, considering the range of densities which is of interest. For example, for a hydrogenic plasma density of  $10^{17} \text{ cm}^{-3}$  (about the upper limit for most high-temperature plasma studies), and a mean ion energy of 10 ev,  $d = 1/n\sigma_d \cong 1.6 \times 10^{-4} \text{ cm}$ , whereas at  $n = 10^{12} \text{ cm}^{-3}$  and 1 kev,  $d \cong 16 \text{ km}$ . Clearly the physical behavior of the plasma can be expected to be quite different over ranges of temperature and density which may be encountered in even a single experiment. See SCATTERING EXPERIMENTS, ATOMIC AND MOLECULAR.

**Relaxation time.** The collision processes in a plasma serve to define the rate of randomization of the energy and direction of motion of the plasma particles. One reason that these processes are important is that they act in opposition to magnetic confinement (especially in the mirror machine) and set an upper limit on its duration. From fundamental work by S. Chandrasekhar and Spitzer, the so-called "relaxation time" for a large angular deflection or for energy exchange comparable with the original energy of a given ion (or electron) colliding with particles of the same kind can be calculated. To a sufficient approximation, this time is given by

$$\tau = 5.7 \times 10^5 (A^{1/2}/Z) (W^{3/2}/n) \quad \text{sec} \quad (36)$$

where  $W$  is in ev.

Returning to the numerical examples given, in this case one has, for protons ( $A = Z = 1$ ), at a density of  $10^{17} \text{ cm}^{-3}$  and an energy of 10 ev,  $\tau = 1.8 \times 10^{-10} \text{ sec}$ ; while for electrons ( $A = 1/1836$ ),  $\tau = 4 \times 10^{-11} \text{ sec}$ , which is a very short time indeed. On the other hand, at  $n = 10^{12} \text{ cm}^{-3}$  and energy of 1 kev, these times become equal to 18 msec and 0.42 msec respectively. It is quite clear that in the first example in the times are small even compared with the cyclotron frequencies of the plasma particles (normally the shortest characteristic times in a magnetically confined plasma) so that heating by magnetic compression would be slight and magnetic confinement would be essen-

tially inoperative, the plasma diffusing like an ordinary gas. On the other hand, in the second example, the times are very long compared with cyclotron periods for reasonable magnetic field strengths, and in slowly varying fields the adiabatic assumptions would be well satisfied. Further increase in the temperature or magnetic field strength would make the conditions even more accurately satisfied.

**Degrees of freedom.** Another interesting consequence of plasma situations where the collision times are long is that for operations performed on the plasma in times short compared with the collision times, the degrees of freedom of the plasma motion are essentially uncoupled. That is to say, compressions perpendicular to the field direction involve only the rotational degrees of freedom, so that the plasma acts as a two-dimensional gas. Similarly, compression along the field lines results in the plasma acting as a one-dimensional gas. The laws of thermodynamics relate the value of the adiabatic gas constant  $\gamma$  to the number of degrees of freedom  $f$ ;  $\gamma = (2 + f)/f$ . Thus for compression transverse to  $B$ ,  $f = 2$ , so that  $\gamma = 2$ . For longitudinal compression  $\gamma = 3$ . The same general thermodynamics laws show that the temperature-vs-particle density law for an adiabatic compression of a gas is that

$$T \propto n^{\gamma-1} \quad (37)$$

Thus for transverse compression  $T \propto n$ , but for longitudinal compression  $T \propto n^2$ ; that is,  $T$  varies as the square of the compression ratio. But if the compression were carried out in a time long compared with collision times,  $f = 3$ ,  $\gamma = 5/3$  (all degrees of freedom coupled), and  $T \propto n^{2/3}$ , a substantially less pronounced variation with density.

**Dynamical friction.** While collisions between an energetic ion and the electrons of a plasma lead to little deflection of the ion's path, owing to the small mass of the electrons, another important effect, dynamical friction, can arise. Provided only that the mean energy of the ion is greater than that of the electrons, the statistically averaged effect of collisions with electrons within the Debye length leads to a frictional drag on the ion which will eventually reduce its mean energy to that of the electrons. This can be an important effect, leading to a damping of the energy of energetic ions immersed in a cold plasma. On the other hand, if the mean ion energy is lower than that of the electrons, the electron collisions will lead to an increase of the ion energy by heating. This latter effect is important in cases where plasma heating is accomplished via heating the electrons first, as in joule or resistive heating.

A single expression can be derived which represents both these effects. This is

$$\left(\frac{dW}{dt}\right)_{ei} = 4\pi\sqrt{2} \ln \Lambda \frac{n_e Z^4 e^4}{(\pi m k T_e)^{1/2}} \frac{m}{M} \left(1 - \frac{W}{3kT_e/2}\right) \quad (38)$$

For ion energies small compared with the mean electron energy, the rate of ion heating ( $T_e$  in ev)

is given by

$$\left(\frac{dW}{dt}\right)_{ei} = 8.8 \times 10^{-8} \frac{Z^2 n_e}{A T_e^{1/2}} \text{ ev/sec} \quad W \ll 3kT_e/2 \quad (39)$$

When the ion energy is substantially greater than the mean electron energy, the general expression takes the form

$$\frac{1}{W} \frac{dW}{dt} = -5.7 \times 10^{-8} \frac{n_e Z^2 / A}{T_e^{3/2}} \text{ sec}^{-1} \quad W \gg T_e \quad (40)$$

This can be integrated to yield an expression for the exponential decay of ion energy as a function of time

$$W = W_0 e^{-t/t_e} \quad (41)$$

$$\text{where } t_e = 1.8 \times 10^7 (T_e^{3/2} / n_e) (A / Z^2) \quad (42)$$

If the electron temperature is low,  $t_e$  may be quite small compared with the mean ion-ion collision time, and therefore will dominate the energy exchange times for high-energy ions. For example, if  $n_e = 10^{14} \text{ cm}^{-3}$  and  $T_e = 10 \text{ ev}$ , then for energetic protons,  $t_e = 5.5 \mu\text{sec}$ . By contrast, if the proton energy is, for example, 10 kev, the corresponding ion-ion scattering time  $\tau$  calculated from Eq. (36) is roughly 6 msec or 1000 times larger. At an electron temperature of 1 kev, however, the two times become about equal.

The expression for  $t_e$  holds as long as the ion velocity is less than the mean velocity of the electrons, that is, as long as

$$W < (M/m)^{1/2} kT_e \quad (43)$$

When this is not satisfied, the dynamical friction rate is somewhat smaller than that just predicted. For protons, this occurs when  $W = 2760 kT_e$ , that is, only at very high energies.

The entire question of energy transfer rates between ions and electrons is one of great importance in plasma research, since in many cases these rates are critical in determining the relative importance of other processes, such as the rate of loss of energy from the plasma by radiation (which arises from the electrons). But as of this writing almost no experimental data exist on these fundamental processes.

## RADIATION FROM PLASMAS

Radiation provides a direct cooling mechanism for a high-temperature plasma. Fortunately indeed for the future of plasma research, theory shows that at even the highest particle densities used in the laboratory, the radiation rates from a plasma are much less than the Planck or black-body value. For example, at a radiation temperature of  $10^8 \text{ }^\circ\text{K}$ , the Planck radiation, being proportional to  $T^4$ , would amount to the almost inconceivable value of  $6 \times 10^{20} \text{ watts/cm}^2$  (see HEAT RADIATION). But the fortunate fact that a tenuous plasma is optically very "thin" over almost all of its emission spectrum



means that, as one might expect from Kirchhoff's law, radiation is greatly reduced compared with the Planck value, so that under the proper circumstances a plasma with a kinetic temperature of  $10^6$  °K might radiate at a rate equivalent to the radiation rate from a black body at radiation temperatures of only a few hundred degrees K. Nevertheless, in many experiments great care must be taken to avoid certain impurity problems so as to keep the radiation from rising to a value where it would overwhelm the means for heating the plasma and keeping it hot.

**Common mechanisms.** Considering collisional processes, there exist three important mechanisms for radiation from a plasma.

The first is the generation of x-rays (bremsstrahlung), which occurs when the plasma electrons are deflected by encounters with the ions (see BREMSSTRAHLUNG). The second mechanism is a similar radiation which occurs when electron-electron collisions occur. This process is important only at very high electron temperature, where the electron motion becomes relativistic ( $T_e$  of order  $mc^2 = 511$  kev). The third mechanism which can occur, and one which under some circumstances may overwhelmingly dominate the radiation losses, is that process which might be called excitation radiation, radiation resulting from the collision of electrons with partially stripped ions (ions with remaining bound electrons) with the production of excited states of the bound electrons and subsequent radiation. Electron-ion bremsstrahlung can be calculated by the methods of quantum mechanics, from which one finds, for a maxwellian distribution, the approximate expression for the radiation per unit volume

$$p_{ei} = 1.4 \times 10^{-34} n_e n_i Z^2 T_e^{1/2} \text{ watt/cm}^3 \quad (44)$$

Here  $Z$  is the ionic charge, and  $T_e$  is measured in degrees Kelvin kinetic temperature. Except at high densities, the radiation rate is nominal, being only some 14 milliwatts/cm<sup>3</sup> for a hydrogenic plasma at  $10^6$  °K and a density of  $10^{14}$ /cm<sup>3</sup> ( $Z = 1$ ,  $n_e = n_i$ ). At  $n = 10^{17}$ /cm<sup>3</sup> this would be  $10^6$  times larger, reaching the respectable figure of 14 kilowatts/cm<sup>3</sup>. Also for a plasma composed entirely of higher  $Z$  ions, since electrical neutrality of the plasma requires that  $n_e = Zn_i$ , the bremsstrahlung radiation rate varies as  $Z^3$  and would therefore be much larger.

Excitation radiation rates can be calculated if the degree of stripping of the ions in the plasma is known. In the past it has been customary to assume that by the time kinetic temperatures of  $10^5$ – $10^6$  °K or higher have been reached, all atoms in a plasma will have been completely stripped of their bound electrons. In this case, there would, of course, be no excitation radiation emitted. The assumption of complete stripping would be valid for a hydrogenic plasma, whose atoms have only a single, easily stripped electron. Unfortunately, it is not possible to create in the laboratory an absolutely pure hydrogen plasma; there will always ex-

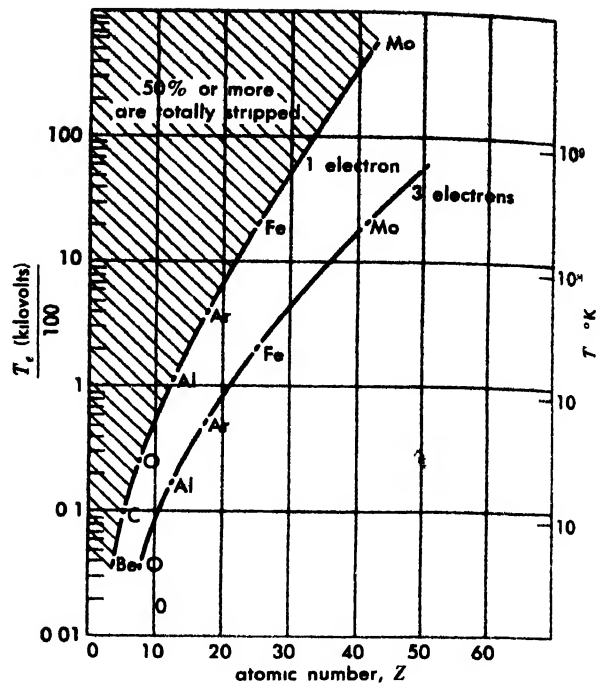


Fig 7 Stripping curve

ist some appreciable number of higher  $Z$  contaminant atoms, such as oxygen.

By means of calculations similar to those employed by astrophysicists in calculating the radiation from the sun's corona, one can compute the degree of stripping of impurity atoms immersed in a low  $Z$  plasma. It is found that the degree of stripping is much less than one would have intuitively assumed. As a result, the role of excitation radiation can be very important, even at quite high kinetic temperatures.

In the calculations it is found that the most interesting and hardy ions are those which are stripped down to one to three or four remaining electrons. Using approximate ionization cross sections, a stripping curve has been calculated for the expected relative abundances of one-electron (hydrogenlike) and three-electron (lithiumlike) ions as a function of atomic number and kinetic temperature (Fig. 7). The figure shows loci for the curve dividing the region of temperature below which a given atom has one or more bound electrons, and for a similar curve for three or more bound electrons. These curves represent steady state values of the stripping. It will be seen that high  $Z$  impurity atoms become completely stripped only at very high temperatures.

Using calculations of this kind, one can determine the expected rates of excitation radiation, per impurity atom, as a function of kinetic temperature. This should be done for all existing states to obtain the total radiation rate. However, it is sufficiently informative to present the results for the two states (one electron and three electrons) just given. The results of these calculations, using approximate excitation cross sections, are presented in Fig. 8. The results are normalized on an atom

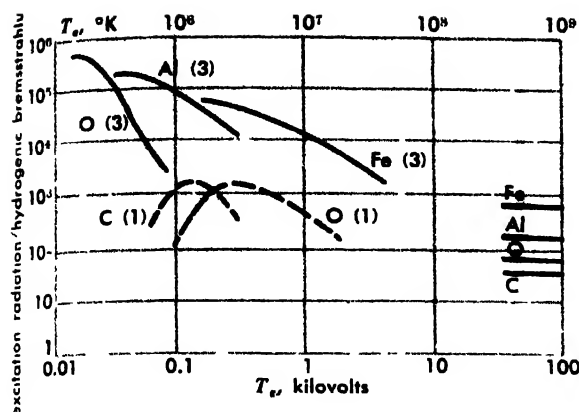


Fig 8 Ratio of two important plasma radiation mechanisms versus temperature for iron, aluminum, oxygen, and carbon.

for-atom basis, against the ordinary hydrogenic bremsstrahlung rate. The feature that is immediately apparent from the curves is that at temperatures of order  $10^6$  °K, the excitation radiation rate is enormous compared with ordinary bremsstrahlung. This radiation is all emitted in the vacuum ultraviolet region of the spectrum:  $\lambda \approx 100$ – $1000$  Å. Only as the temperature reaches  $10^7$  °K or more does it drop to more reasonable values. This shows that at the lower temperatures excitation radiation may provide a very rapid cooling process for a hydrogenic plasma, even with relatively small percentages of impurities. Secondly, it is apparent that once very high temperatures are reached, the rate falls to about the ordinary bremsstrahlung values (indicated by the marked lines along the right edge of the plot).

The preceding calculations represent steady-state rates. However, in experimental plasmas the duration of the experiment may be too short for the plasma ions to reach the steady-state degree of stripping. In this case, calculations show that the radiation rate may be substantially higher than that just indicated.

At relatively low plasma temperatures and very high impurity densities, self-absorption of the radiation may become important and reduce the calculated loss somewhat. This usually occurs only at a very high absolute radiation flux, however, since the onset of self-absorption signals the approach to equilibrium or Planck radiation levels.

**Other mechanisms.** The mechanisms described thus far produce their main radiation flux in the x-ray and the vacuum ultraviolet part of the spectrum. Going toward longer wavelengths, the next significant radiation region is not reached until the long wave infrared or the short wavelength microwave region is reached. In this region, the plasma again possesses mechanisms which can produce appreciable radiation fluxes. Being limited by the Rayleigh-Jeans value (which varies as the square of frequency), this long-wavelength radiation does not comprise an appreciable energy loss mechanism except at the highest electron temperatures. The ra-

diation in this region is important only for a plasma immersed in a strong magnetic field. It arises simply from the centrifugal acceleration of the electrons as they move in helical orbits in the magnetic field. Measurement of this radiation provides another possible way to determine the electron temperature of the plasma, by applying techniques similar to those employed in radio astronomy.

[R.F.P.]

**Bibliography:** H. Alfvén, *Cosmical Electrodynamics*, 1950; S. C. Brown, *Basic Data of Plasma Physics*, 1959; R. F. Post, Controlled fusion research—an application of the physics of high temperature plasmas, *Revs. Modern Phys.*, 28:338–362, 1956; R. F. Post, Fusion power, *Sci. American*, 197(6):73–84, 1957; L. Spitzer, Jr., *Physics of Fully Ionized Gases*, 1956; United Nations, *Proc. Second Intern. Conf. Peaceful Uses Atomic Energy*, vols. 31–32, 1958.

## Plasmal

Those aldehydic components of lipids which give positive color tests with reagents used for detecting aldehydes in tissues. The Feulgen test for plasmalogens depends on the liberation of plasmal, principally palmitaldehyde and stearaldehyde, from these lipids by the action of mercuric chloride and acetic acid. See LIPID; PHOSPHATIDE.

[H.E.C.; R.H.C.]

## Plasmodroma

In some taxonomic systems this is a subphylum of Protozoa, including the Mastigophora, Sarcodina, and Sporozoa. The Ciliata are excluded. Locomotor organelles, if present, may be flagella or pseudopodia, but not cilia. There is no nuclear dimorphism like the micronucleus and macronucleus as seen in typical ciliates. The subphylum Plasmodroma is not recognized as a subdivision of Protozoa by certain modern taxonomists.

[R.P.H.]

## Plaster

A plastic mixture of solids and water which sets to a hard, coherent solid and which is used to line the interiors of buildings. A similar material of different composition, used to line the exteriors of buildings, is known as stucco. The term plaster is also used in the industry to designate plaster of paris.

Plaster is usually applied in one or more base (rough or scratch) coats up to  $\frac{3}{4}$  in. thick, and also in a smooth, white, finish coat about  $\frac{1}{16}$  in. thick. The solids in the base coats are hydrated (or slaked) lime, sand, fiber or hair (for bonding), and portland cement (the last may be omitted in some plasters). The finish coat consists of hydrated lime and gypsum plaster (in addition to the water). See LIME (INDUSTRIAL); MORTAR; PLASTER OF PARIS.

[M.C.M.]

## Plaster of paris

The hemihydrate of calcium sulfate,  $(\text{CaSO}_4 \cdot \frac{1}{2}\text{H}_2\text{O})$  made by calcining the mineral gypsum  $(\text{CaSO}_4 \cdot 2\text{H}_2\text{O})$  at temperatures up to  $250^\circ\text{C}$ . It is used for making plasters, molds, and models. For a

description of the casting method of ceramic forming, see CERAMIC TECHNOLOGY.

When the powdered hemihydrate is mixed with water to form a paste or slurry, the calcining reaction is reversed and a solid mass of interlocking gypsum crystals with moderate strength is formed. Upon setting there is very little, if any, dimensional change, making the material suitable for accurate molds and models.

The chemical reaction of hydration requires 18.6 lb of water for each 100 lb of plaster of paris; any excess of water over this makes the mixture more fluid and it eventually evaporates, leaving a porous structure. In general, the greater the porosity, the lower the strength of the set plaster. Plaster of paris molds for ceramic casting must have a certain minimum porosity in order to absorb water from the slip; thus, for this application, the amount of water added represents a compromise between the conflicting properties of high strength and high porosity.

Various types of plaster, varying in the time taken to set, the amount of water needed to make a pourable slip, and the final hardness, are made for different applications. These characteristics are controlled by the calcination conditions (temperature and pressure) and by additions to the plaster. For example, hydrated calcium sulfate ( $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ ) greatly accelerates the setting time and therefore the use of utensils contaminated with set plaster is to be avoided. See GYPSUM; PLASTER.

[M.C.M.]

## Plastic deformation of metal

The permanent change in shape of a metal as a result of exceeding the elastic limit. The plasticity of a metal permits it to be molded or pressed, without rupture, into various forms that are retained after the pressure of molding or pressing has been removed. For a discussion of conditions under which changes in shape are not permanent, see ELASTICITY.

The properties of metals that are associated with plastic deformation are ductility (the ability of a metal to be deformed considerably before breaking), creep (the time-dependent deformation of metal under stress), and malleability (the ductility of a metal under the particular conditions of a given metal-forming operation). See METAL; METAL, MECHANICAL PROPERTIES OF; METAL FORMING.

**Ductility.** A substance is brittle if it breaks without deforming; it is ductile if it can deform considerably before breaking. Ductility is important in metals for two reasons. It permits a metal to be put into commercially desirable shapes without breaking, and it permits a metal to absorb shocks and blows in service that could break a stronger but more brittle material. One simple measure of ductility is the reduction in cross-sectional area that a metal sample will show when pulled apart in simple uniaxial tension. There are metals that are as brittle as glass, showing virtually

no reduction in cross-sectional area in a tensile test, whereas others will be as ductile as taffy, stretching out in a tensile test until the cross section of the sample has contracted to a mere filament. Ductility is not a fixed property of a given metal but will depend upon such factors as the temperature, the speed with which the metal is broken, the size and shape of the metal, impurities that may be present in the metal, the environment in which the metal finds itself, and the manner in which the metal deforms prior to breaking.

In a tensile test, the metals crystallizing in the face-centered cubic system (for example, copper, aluminum, nickel, lead) are generally ductile at all temperatures and rates of deformation strain although even these metals (especially those with large grain size) show some brittleness in some combinations of temperatures and deformation rates, as shown by the valley A-A in Fig. 1. Since the metal deforms plastically at temperatures below this valley and viscously at temperatures above it, this brittleness is usually associated with the high degree of deformation occurring in the few regions of the internal structure of the metal where viscous flow first makes its appearance. The term equicohesive brittleness is usually applied to this phenomenon. A quantitative measure of ductility is the ratio of the cross-sectional areas of the test piece before ( $a_0$ ) and after ( $a_f$ ) the piece failed. Since the range of this ratio is large the logarithm of the ratio is usually plotted (see Figs. 1, 2, 3, 4) and for theoretical reasons the natural logarithm may be used.

Certain of the close-packed hexagonal metals (notably zinc and cadmium, but not titanium or zirconium) and certain of the body-centered cubic metals (notably iron, chromium, tungsten, and molybdenum, but not vanadium or tantalum) become brittle at low temperatures and at high deformation rates (Fig. 2). This phenomenon is the basic cause for the dramatic failure of steel structures in cold weather under shock (for example, the breaking in half of the ocean freighter *Flying En-*

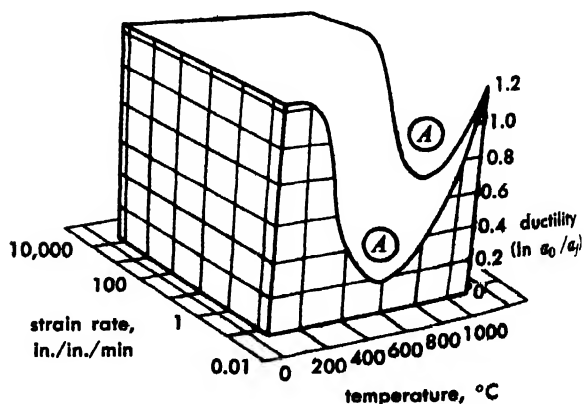


Fig. 1. Ductility of copper in a tensile test (measured by the natural logarithm of the ratio of cross-sectional area of tensile specimen before ( $a_0$ ) and after ( $a_f$ ) the test) as a function of temperature and test

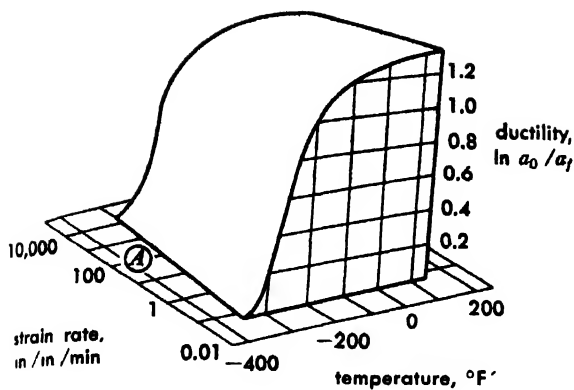


Fig 2 Ductility of annealed mild steel in a tensile test as a function of temperature and test speed. Metal is virtually glass-brittle at low temperatures and high strain rates (region A). (From T. Toh and W. M. Baldwin, Jr., in W. D. Robertson, ed., *Stress Corrosion Cracking and Embrittlement*, Wiley, 1956)

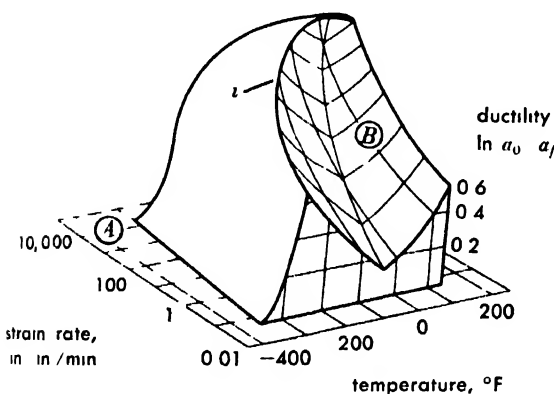


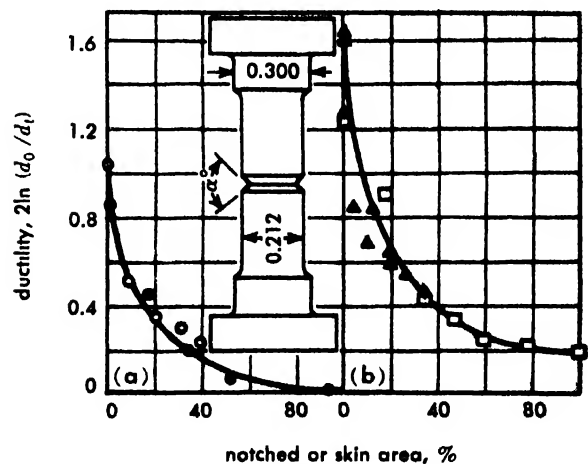
Fig 3 Ductility of annealed mild steel charged with hydrogen as a function of temperature and test speed. Presence of hydrogen has depressed the ductility of the steel in region B. Brittleness in region A exists whether hydrogen is present or not. (From T. Toh and W. M. Baldwin, Jr., in W. D. Robertson, ed., *Stress Corrosion Cracking and Embrittlement*, Wiley, 1956)

surprise in the North Atlantic in January, 1955).

Dissolved impurities which diffuse through, or show some mobility in, the host metal will affect the ductility of the host metal in certain ranges of temperature and deformation rates. If, by a suitable choice of temperature the mobility of the impurity can be changed so that its value is in some critical proportion to the deformation rate, the movement of the host metal atoms going to make up the deformation mechanism may well be hampered and brittleness will result. Figure 3 shows, for example, how steels containing hydrogen display a brittleness over certain temperatures and deformation ranges (region B of Fig. 3). This is in addition to the low-temperature brittleness which steels possess, whether they contain hydrogen or not (region A of Figs. 2 and 3).

Notches on the surface of or holes within metal specimens under tension will affect their ductility

in two ways. They will lower the general level of ductility of all metals and they will shift the cliff which bounds the low-temperature brittle range of Fig. 2 to higher temperatures and lower strain rates. Both of these effects become greater as the area of cross section occupied by the notches or holes becomes larger (Fig. 4); and both of these effects (for a given proportion of the cross section occupied by notches or holes) will vary in intensity from metal to metal. These considerations are important because a given steel, when tested as a smooth bar in a simple tension test, may be quite ductile, and yet when fabricated into a machine part containing notches or holes may be dangerously brittle. The embrittling effects of blow-holes from improper casting or voids introduced by improper annealing (for example, the hydrogen-annealing of oxygen-bearing copper) stem from this same cause. Undissolved impurities whose hardnesses are markedly different from that of the host metal (either harder or softer) are embrittling because they effectively behave as holes (discontinuities) in the matrix. In a similar vein, brittle skins put on metals intentionally (carburized or nitrided cases put on steel, or chromium and copper plates) or unintentionally (oxygen- or nitrogen-rich surfaces in titanium annealed in air) will embrittle the base metal. At the slightest deformation



○ Notch tensile data for SAE 1020 steel (circles represent ductility with 60° notches. Upper end of vertical line represents ductility with 90° notches. Lower end represents ductility with 30° notches).

● Data for carburized SAE 1020 steel.

▲ notch tensile test data for copper.

■ data for hydrogen-embrittled copper.

Fig. 4. Ductility of (a) annealed mild steel and (b) copper as a function of the percentage of cross-sectional area occupied by a notch of the type shown in the sketch (open symbols) or as a function of the percentage of cross-sectional area occupied by a brittle skin (filled symbols). (From G. W. Form and W. M. Baldwin, Jr., *The Effect of Brittle Skins on the Ductility of Metals*, Am. Soc. Testing Materials, Preprint 79, 1956)

## Plastic deformation of metal

the brittle skin cracks, and from this point on, the base metal behaves as a notched specimen, as shown in Fig. 4.

**Creep.** Creep is the continuing, or time-dependent, deformation that a metal or any substance evidences when put under stress. It is unlike plastic deformation, which is fixed (unchanging with time) under stress and permanent (remaining after the stress is removed), and unlike elastic deformation, which is fixed and transient (being recovered after the stress is removed). Creep or time-dependent deformation is the predominant reaction of metals to stress when temperatures are high, that is, in the upper half of the temperature range from absolute zero to the melting point of the metal, and when stresses are low. Elastic deformation predominates at low temperatures and low stresses; plastic deformation predominates at high stresses. As the temperature approaches the melting point of the metal, creep tends toward a truly viscous phenomenon since the deformation rate tends to become constant under a constant stress. The higher the applied stress  $s$  the greater the deformation rate  $\dot{\epsilon}$  will be according to the formula  $\dot{\epsilon} = As^n$  where  $A$  and  $n$  are constants,  $n$  being far in excess of unity. At lower temperatures creep becomes more complex. After an immediate elastic or plastic response to an applied load, the metal will deform with time at a rate that slowly falls off. The deformation may settle down to what appears to be a steady rate, but ultimately, even in tests where the stress is held constant, will accelerate at an ever faster pace until the metal breaks. The terms primary, secondary, and tertiary creep are used to describe the stages when the creep rate is falling off, remaining constant, or accelerating, respectively. The accelera-

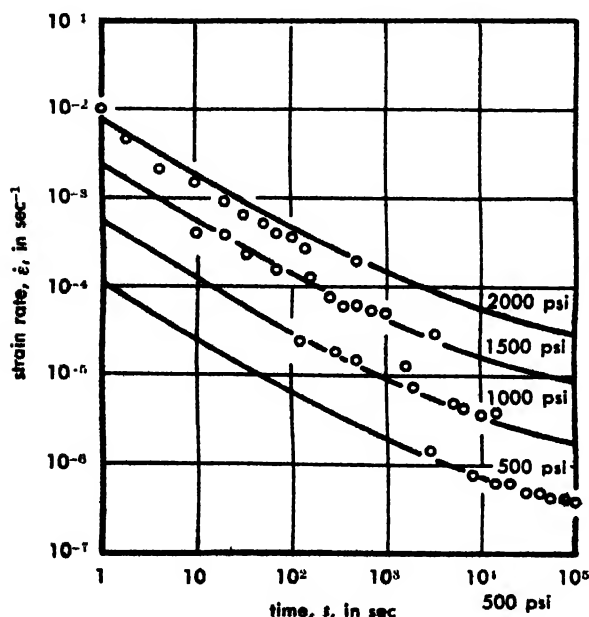


Fig. 5. Creep rate of aluminum for four different stresses plotted against time (temperature being constant at 277°C).

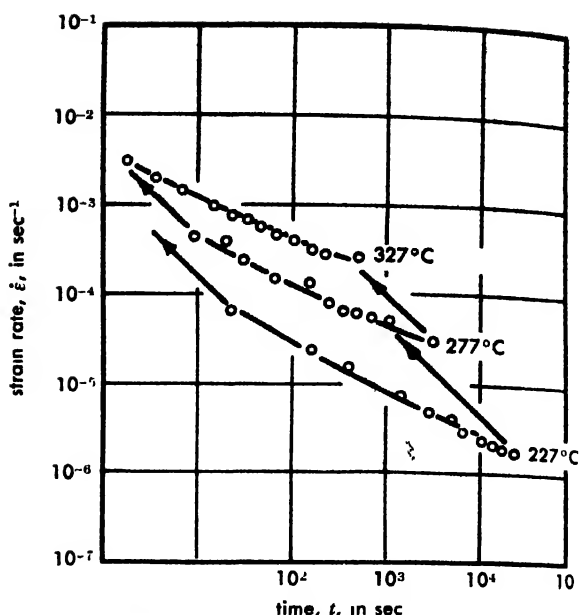


Fig. 6. Creep rate of aluminum at three different temperatures plotted against time (stress being constant at 1500 psi).

tion of creep rate during tertiary creep is usually ascribed to the breakdown of the internal metal structure as a result of the deformation which has been going on. This breakdown can usually be seen under the microscope as voids or cracks opening up at grain boundaries or junctures of grain boundaries. Inseparable from the creep phenomenon is creep recovery. A metal that crept under load will recover some of the deformation it so accumulated when unloaded. Part of the recovery will be immediate, being purely elastic in nature, but part will slowly accrue with time at an ever-decreasing rate.

Although the creep rate varies with time in the more complex forms of creep (at temperatures not close to the melting point of the metal) it can be predicted with fair success for a given temperature or stress if it is known over a given period of time at another temperature and stress. Figure 5, for example, shows the logarithm of the creep rate of aluminum plotted against the logarithm of time for four different applied stresses, the temperature being held constant. It is obvious that increasing the stress merely shifts the curve upward without distorting it in any way. Similarly Fig. 6 shows the logarithm of the creep rate of aluminum plotted against the logarithm of time for three different temperatures, the stress being held constant. Increasing the temperature in two 50° steps shifts the curve bodily upward and to the left in equal amounts. These relations are expressed by the formula

$$\dot{\epsilon} = s^n f \left( \frac{10^3}{T - T_0} \right)$$

Here  $\dot{\epsilon}$  is the strain rate in in./in./sec,  $s$  is the applied stress,  $n$  is a constant (equal to 3 in the case of aluminum taken for illustration),  $f$  is the func-

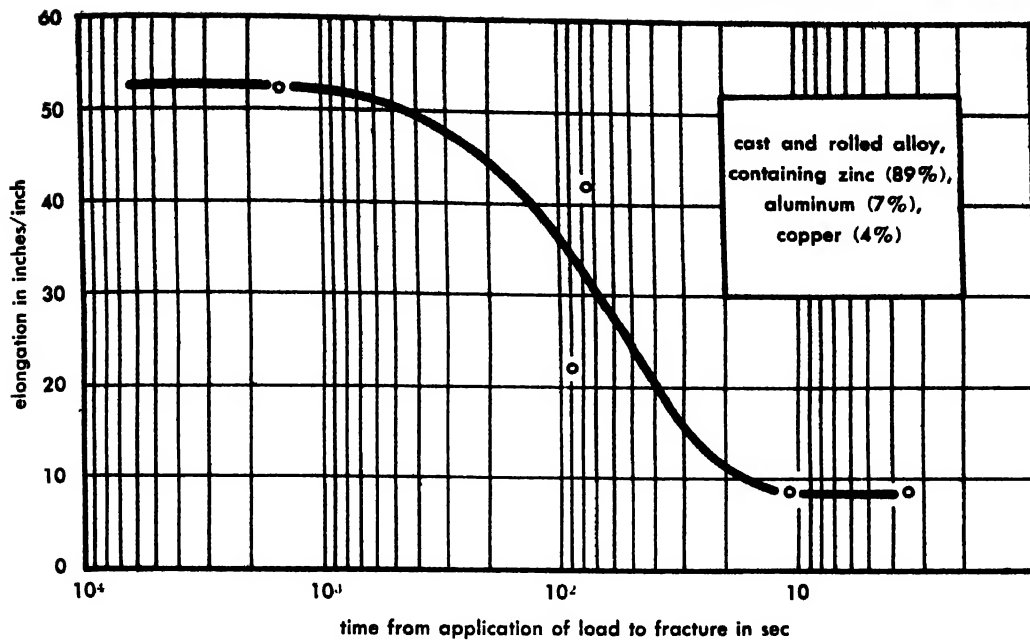


Fig. 7. The ductility of a cast and rolled zinc-aluminum-copper eutectic alloy, containing 89% zinc, 7% aluminum, and 4% copper (here measured as elongation in a tensile test), drops off as the test is speeded up. W. Rosenhain, J. L. Haughton, and K. E. Bingham who reported these data noted that

the alloy "flew to pieces under the shears unless cut slowly" and that the alloy "broke like glass when bent rapidly or struck, but could be doubled on itself if bent slowly." (From W. Rosenhain, J. L. Haughton, and K. E. Bingham, *Zinc alloys with aluminum and copper*, *J. Inst. Metals*, 23:261-324, 1920)

tional relation between creep rate  $\dot{\epsilon}$  and time  $t$ ,  $R$  is the gas constant,  $T$  is the absolute temperature, and  $Q$  is a constant equal to the activation energy for diffusion of the metal in question. The strain rate will vary with each metal and depends upon factors such as temper and grain size.

The ductility of creeping metals falls off as the deformation rate is speeded up (Fig. 7). This behavior is found in viscous or visco-elastic materials in general, for example, in crazy putty, in suspensions of sand in olive oil, and in glass. The strain rate  $\dot{\epsilon}_c$  at which the ductility suddenly drops, moves to higher values as temperature is raised in accordance with an Arrhenius type of equation,

$$\dot{\epsilon}_c = Ae^{-Q/RT}$$

where  $A$ ,  $Q$ , and  $R$  are constants.

**Malleability.** In the strictest sense this is the ability of a metal to be rolled or hammered without breaking, although it frequently is interpreted in a broader sense as the ability of a metal to be formed by any mechanical process. It is thus a very practical measure of the ductility of a metal under the particular conditions of a given forming operation. Since metals are formed in commercial operations at elevated temperatures as well as at room temperature, metal producers will consider both the hot and cold malleability of a metal. The terms hot-short or cold-short refer to metals lacking hot or cold malleability.

Any of the forms of brittleness discussed in the paragraph on ductility may be responsible for some form of hot- or cold-shortness. Most metals that are

commercially formed are free from most types of brittleness, and the cases of hot- or cold-shortness to which they may occasionally succumb are caused by an unwanted appearance of impurities.

Hot-shortness usually is caused by some second phase (frequently present in fractions of a per cent) that turns liquid at the temperatures in question and which, by running between the grain boundaries, destroys all cohesion of the host metal. The relative surface tensions of the liquid phase and the host metal will determine whether the liquid phase will run freely or will ball up in harmless droplets.

Lead can render copper hot-short if present in quantities more than 0.02%. Brasses (copper-zinc alloys) can tolerate no more lead than copper itself, at least in the so-called alpha range containing 0-35% zinc in which the crystal structure of the brass is the same as that of copper (face-centered cubic crystal structure, *see CRYSTAL STRUCTURE*). If enough zinc (35-45%) is present in brass to produce the beta crystal structure (body-centered cubic crystal structure), the surface tension relationships between brass and lead are so radically altered that the lead now balls up harmlessly, and such brasses can tolerate 2-3% lead, or more, and still be hot-malleable.

Hot-shortness caused by liquid phases can be counteracted by the addition of third elements which combine with the liquid phases either to form solids or to disturb the surface tension relationships between the liquid and the host metal. As an example, the tolerance of copper for lead increases sharply as oxygen is added to copper (Fig. 8). A



### Impurities commonly rendering some metals hot-short and their counteractants

Host metal	Impurities producing hot-shortness	Counteractant
Iron	Sulfur (0.017)*	Manganese
Austenitic stainless steels	Lead (0.005)	Oxygen
Nickel	Sulfur (0.004)	Magnesium, manganese
	Silicon (0.3)	Manganese
	Lead (0.004)	Oxygen
Cobalt	Sulfur (0.010)	Magnesium
Copper	Lead (0.020)	Oxygen
	Bismuth (0.002)	Oxygen

\* The numbers in parentheses are the critical limits of the impurity (%) which the host metal can tolerate for hot-rolling.

brief résumé of some common impurities that render metals hot-short and some counteractants to these impurities are assembled in the table.

If impurities render a metal cold-short it is usually because they occur as fine precipitates either at the grain boundaries of the host metal or throughout the entire matrix. As noted above, if the hardness of these particles is sufficiently different from that of the matrix they act effectively as holes and embrittle the metal through a notch effect. Fig.

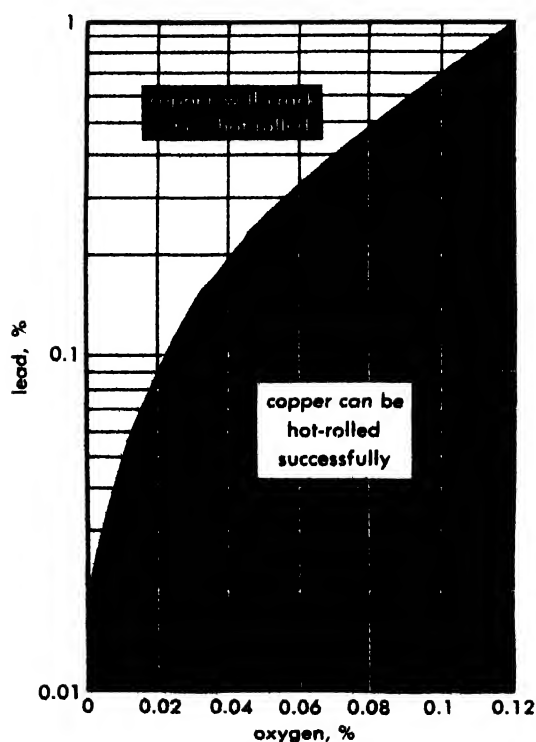


Fig. 8. This chart shows the lead and oxygen contents that copper can tolerate and be hot-malleable. The richer copper is in oxygen, the more lead it can contain and not be hot-short. (From W. M. Baldwin, Jr., *Rolling Copper and Copper Alloys*, in *Nonferrous Rolling Practice*, Am. Inst. Mining Met. Engrs., Inst. Metals Div., Symposium Ser., vol. 2, 1948)

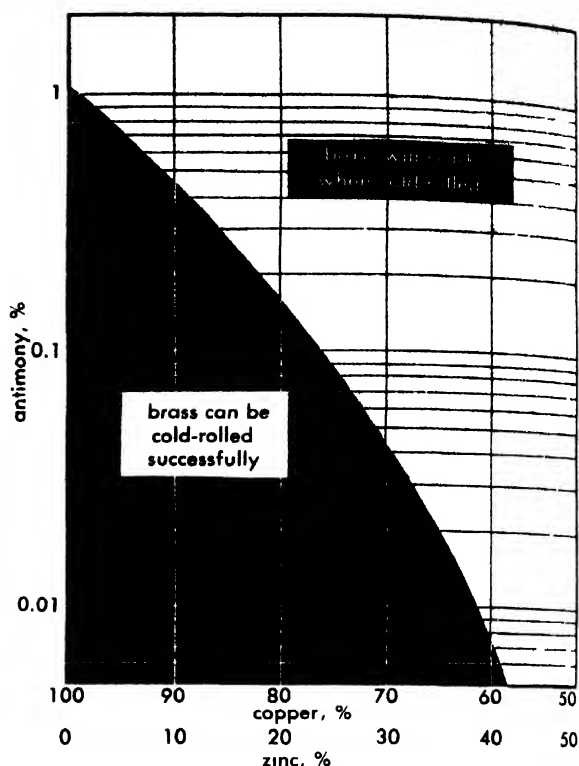


Fig. 9. This chart gives the maximum amount of antimony that brasses of different zinc contents can contain and still be cold-malleable. (From W. M. Baldwin, Jr., *Rolling copper and copper alloys*, in *Nonferrous Rolling Practice*, Am. Inst. Mining Met. Engrs., Inst. Metals Div., Symposium Ser., vol. 2, 1948)

ure 9, for example, gives the maximum amount of relatively hard antimony that may be tolerated in relatively soft brasses of different zinc contents if the metal is to be cold-rolled. These curves lie just above the solubility curve for antimony in brass. Antimony in excess of its solubility limit appears as a hard precipitate in the grain boundaries of the brass and produces cold-shortness. See METALLURGY; STRESS AND STRAIN. [W.M.B.]

**Bibliography:** H. C. H. Carpenter and J. M. Robertson, *Metals*, 2 vols., 1939; L. A. Rotherham *Creep of Metals*, 1951.

### Plasticity

The property of a solid body whereby it undergoes a permanent change in shape or size when subjected to a stress exceeding a particular value, called the yield value. Many solid materials obey Hooke's law at low stresses, but as the stress is increased, departures from Hooke's law occur, and some plastic flow takes place; that is, the material does not completely recover its original shape or size when the stress is released.

Plastic behavior is often accompanied by time-dependent effects such as creep (the increase in strain with time at constant stress), stress relaxation (the decay of stress with time at constant strain), and elastic after-effect or recovery (the gradual decrease to a li permanent strain

when the stress is removed). The study of these phenomena in all their manifestations is the science of rheology. For the influence of dislocations on plastic flow in crystals, see CRYSTAL DEFECTS. See also CREEP OF MATERIALS; ELASTICITY; HOOKE'S LAW; PLASTIC DEFORMATION OF METAL; RHEOLOGY; STRESS AND STRAIN. [R.F.S.H.]

## Plastics fabrication

An operation including a variety of methods for converting fluids, pastes, suspensions, powders, granules, sheets, and special forms into various solid shapes. Many end uses depend on mechanical alterations of an intermediate plastic form, in the same way that a great variety of objects can be made from stock metal shapes. Some plastics are almost as readily machinable as metals. Adhesive bonding can be used in a manner similar to the welding of metals.

The numerous complicated combinations of fabricating methods can be simplified greatly if final finishing operations are not considered, just as the production of standard steel shapes is much less complicated than the forming of finished steel objects. With this simplification, plastics fabrication is used to produce films, coatings, molded and cast objects and standard shapes, such as sheets or rods.

**Films and coatings.** Films and coatings can be produced by identical processes. Films have thicknesses of 0.2–50 mils (0.0002–0.05 in.) and are usually flexible. The term foil may be used for thin, flexible, transparent material used for packaging, whereas sheets or sheeting refers to thicker material that is too rigid for films.

The essential steps in the formation of films and film coatings are preparation of the plastic feed stock (compounding and mixing), formation of the film (drying and conditioning), and finishing. Depending on the actual process, the plastic feed stock may be in the form of lacquers, enamels, suspensions or emulsions, pastes, or plastic masses. The principal methods of film forming are calendering, casting, extrusion, and dip-, knife-, or roll-coating.

**Calendering.** A calender usually consists of four large rolls rotating as shown in Fig. 1. A plastic mass is fed between two rolls which squeeze it out into a film that passes around one or more addi-

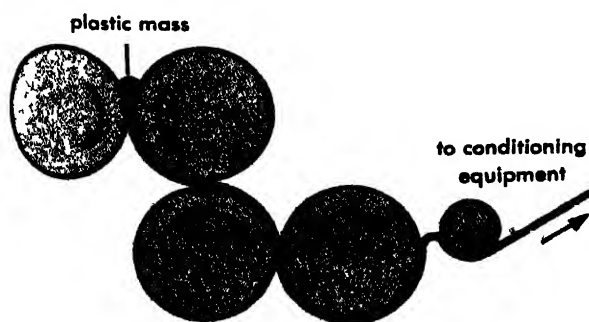


Fig. 1. A four-roll calender.

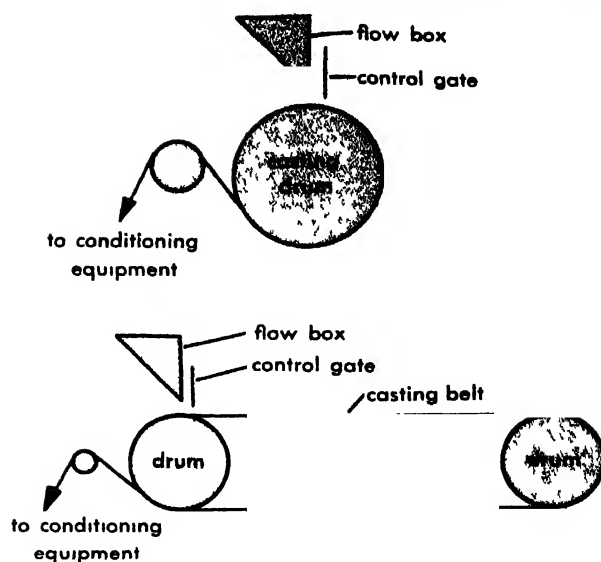


Fig. 2. Casting machines for film formation.

tional rolls before being stripped off as a continuous film. Fabric or paper may be fed through the opening between the last two rolls so that the film is pressed into the surface to produce coatings.

This expensive machine is particularly well suited for the production of large quantities of vinyl films and coatings in thicknesses of 0.003–0.040 in. and widths up to 92 in. It is also used in the rubber industry to produce similar films and coatings.

**Casting.** The technique of casting films on large polished wheels has long been used for photographic film. Casting is also done on polished metal belts or bands. Both methods are shown diagrammatically in Fig. 2. In this method of film formation, a solution of the plastic flows onto a highly polished moving surface where some heat is applied to evaporate part of the solvent, and to produce a film with enough strength to be self-supporting at the point of removal. In wheel casting, wheels or drums from 12 to 18 ft in diameter and up to 6 ft wide are used. The outer surface is carefully ground, polished, and plated with nickel, chromium, or silver. The finished surface must be mirror-smooth because even minute defects are reproduced in the film. The belt or band machine consists of a wide metal belt stretched between two rotating drums housed in a dryer. Widths up to 100 in. are available, and the drums are spaced as much as 300 ft apart.

Thicknesses of cast films vary from 0.0005 to 0.004 in. The wheel process is capable of producing better surfaces and is used for photographic film, but the belt method can be used for faster production. Many plastic films such as cellulose acetate and acetate-butyrate, polyesters, and the vinyls are produced by casting.

Cellophane is also produced by a process called casting. In this case, the film is formed by chemical coagulation. The method of formation is the same as that for viscose rayon except that a wide band

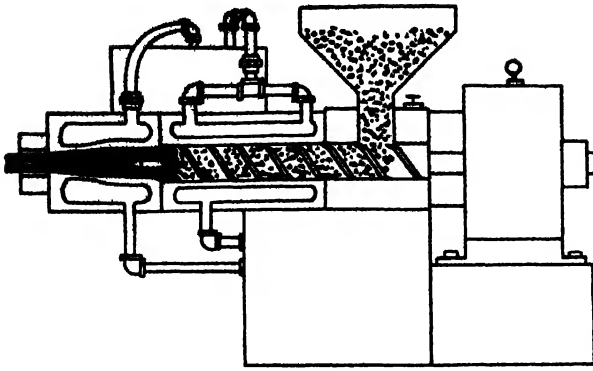


Fig. 3. An extruder for plastics. (From C. C. Winding and R. L. Hasche, *Plastics*, McGraw-Hill, 1947)

of liquid instead of liquid filaments is pumped into the coagulating solution. See FIBER, MAN-MADE.

**Extrusion.** Film is produced by two extrusion methods: the slot-die, or sheet, method and the large-diameter inflated-tube process. An extruder is shown in Fig. 3. In the slot-die method, dies having openings of as little as 0.005 in. and widths up to 72 in. are used. The plastic material emerges from the die in the form of a hot film which must be cooled rapidly. Thin films cannot be made by this method. Thicknesses vary from 0.005 to 0.250 in. As indicated by the thickness range, sheets as well as films can be produced.

Inflated-tube extrusion for film formation involves extruding a thin-walled tube, expanding it while hot, cooling, collapsing, and slitting the tube lengthwise to produce a film with a width equal to the circumference of the inflated tube. The process is indicated diagrammatically in Fig. 4. Inflated diameters from a few inches to 6 ft or more have been used, giving film widths up to 20 ft, much wider than can be made by any other method. Most polyethylene film used in packaging, building construction, and agriculture is made by the inflated-tube process. Thicknesses vary from 0.0005 to 0.004 in.

**Coating processes.** Various coating methods are used to apply plastics to fabrics and papers to produce either finished goods or intermediates for other fabrication processes. The simplest of these is dip coating in which a continuous web passes down into a vat containing a plastic solution, over a submerged roller, and back up out of the solution. In another method, a plastic in solution is poured onto a moving sheet and a knife-spreader distributes it evenly over the sheet. Other methods are similar to printing processes in which one roll rotates in a solution and transfers the plastic material to an intermediate roll which in turn rolls a coating onto paper or fabric.

**Molding.** The molding of small objects is the application normally thought of when plastics are mentioned, although the total tonnage used for films, sheets, and laminates is actually greater. Molding involves filling a mold cavity with a plastic fluidized by heat and pressure, which is allowed

to solidify to produce an object that requires only finishing operations. With the exception of laminating and sheet forming, there are three principal plastic molding processes which convert powders or granules (or simple preforms made by compressing powders) into finished molded objects.

**Compression molding.** In this process, the two halves of molds such as those shown in Fig. 5 are attached to the two platens of a hydraulic press one of which is moved by the ram. The platens are

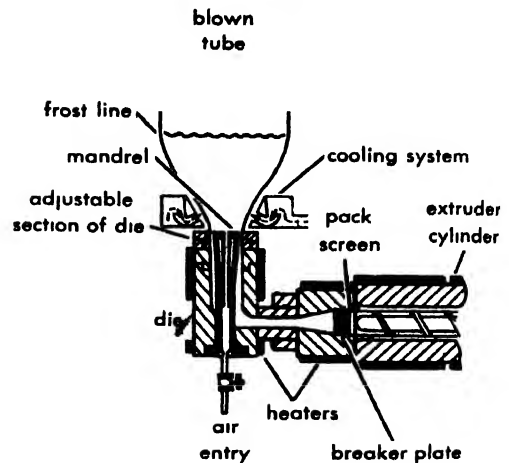
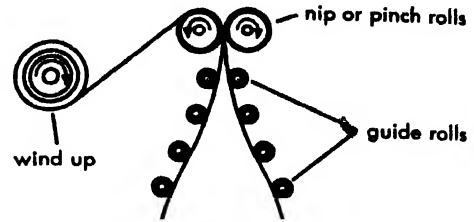


Fig. 4. The inflated-tube process. (U.S.I. Industrial Chemicals Co.)

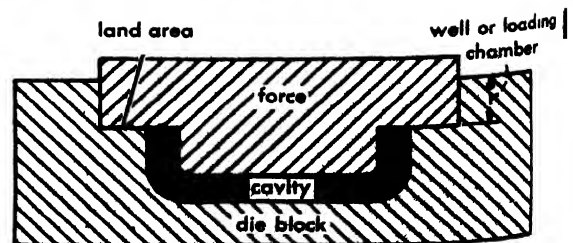
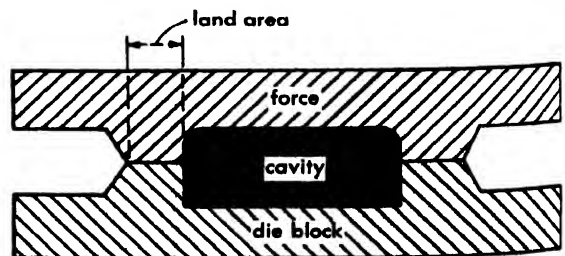


Fig. 5. Two types of compression molds. (From C. C. Winding and R. L. Hasche, *Plastics*, McGraw-Hill, 1947)

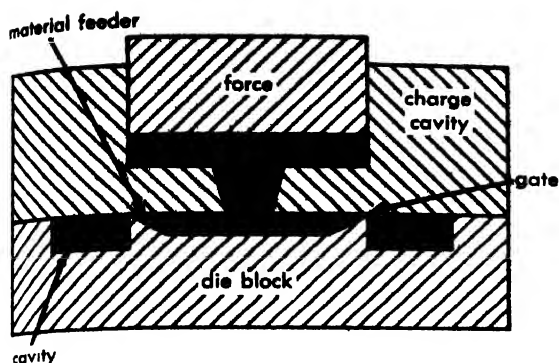


Fig. 6. A transfer mold. (From C. C. Winding and R. L. Hasche, *Plastics*, McGraw-Hill, 1947)

heated with steam or electricity to supply heat to the molds. The cavities are filled with a predetermined amount of powder or preforms, and pressure is applied to bring the two halves of the mold together. The heat and pressure fluidize the plastic, and the mold is closed very slowly until it is seated on the "land" areas.

There are two types of plastics: thermosets and thermoplasts. The former "sets up" or cures with the application of heat; the latter requires cooling for complete solidification. Thermosets can be removed from the mold in a very short length of time. If a thermoplast is used, the mold must be cooled before opening. Compression molding can be used for most plastic materials, but it is most frequently employed in molding phenolics, ureas, melamines, and other thermosets.

**Transfer molding.** This process is similar to compression molding except that fluidization is accomplished in an outside chamber and the fluid is forced into the mold cavities. This is done in molds typified by Fig. 6. The transfer from the charge cavity to the product cavity is accomplished by the closing of the press. The two main halves of the mold come together first; additional movement of the ram pushes the fluid into the charge cavity. Since fluidization takes place outside the mold cavity, thinner sections and more delicate inserts may be used. Some excess plastic is always left in the connecting passages; this must be removed and discarded before the cycle is repeated. This method is normally employed with thermosetting plastics such as the phenolics, ureas, or melamines.

**Injection molding.** In this process, a hot thermoplast is injected into a cooled mold, thereby obviating the need for alternately heating and cooling the mold. Granular plastic material is placed in a hopper from which it is fed in a predetermined quantity to a heated chamber located in front of a hydraulically operated piston. See Fig. 7. After the mold is closed, the plunger forces the plastic through the heated cylinder, then through a nozzle into the sprue in the front half of the mold and on through the runners into the mold cavities. Pressures up to 25,000 psi are used for injection. The plastic cools rapidly and is soon ready for ejection. The entire cyclic operation may be made auto-

matic. From 1 to 300 oz of material may be injected with each shot, depending on the size of the machine. This is the most widely used molding method for thermoplasts. It is capable of high production rates, particularly if multicavity molds are used. Nearly all of the thermoplasts are, or can be, molded by this process.

**Casting.** Plastic materials can be cast both by the use of melts in a manner similar to that used for metals and by polymerizing liquid monomers in a mold. As described previously, the same term is used in film formation. Most plastics are not sufficiently fluidized below decomposition temperatures to fill molds without the application of pressure. However, a few special plastic compounds such as certain phenolics, ethyl cellulose-wax mixtures, and highly plasticized cellulose acetate-butyrate may be melted and poured into molds. These materials are cast into large punches, dies, and blocks used in forming sheet metals.

The casting of acrylates, polystyrene, polyesters, and epoxides depends on polymerization rather than cooling for hardening. Liquid monomers or partially polymerized solutions utilizing a monomer as the solvent are mixed with a catalyst and carefully poured into a mold. Conditions are adjusted so that polymerization takes place relatively slowly, usually at temperatures below 70°C. Since the casting liquid has a relatively low viscosity, mold surfaces are reproduced accurately. If polished molds are used, a high polish is obtained on the finished article. Polished sheets, rods, and tubes of clear polymethyl methacrylate are obtained by using plate glass or stainless steel molds.

Casting is also used for "potting," or encapsulation, of various objects. Biological or other types of specimens may be embedded in clear castings for preservation and display. Entire electrical circuits may be potted by placing them in a container and pouring the casting liquid around them.

**Laminating.** Laminates are plastic shapes containing reinforcing materials such as fibers, cloth, paper, or wood veneers. They vary all the way from reinforced plastics containing less than 10% fibers to plywood which usually contains 8-20% of a plastic bonding agent. The most common laminates are made of several layers of plastic-impregnated cloth or paper bonded together by heating under pressure to fluidize the plastic until it forms a con-

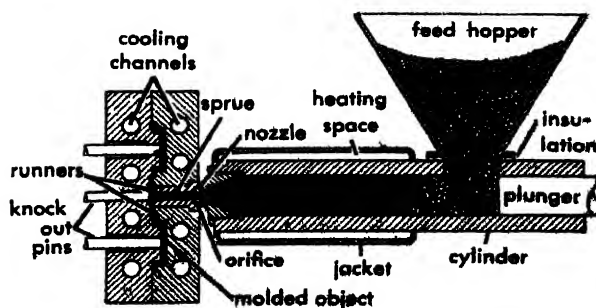


Fig. 7. An injection molding machine. (From C. C. Winding and R. L. Hasche, *Plastics*, McGraw-Hill, 1947)

tinuous phase surrounding the fibers. Mats of fibers may be used in place of cloth or paper.

Both molded and standard shapes may be made by lamination. If a molded shape such as a boat is to be made, a mat of fibers (usually glass), is first laid inside a female mold having the shape of the hull. The fibers are impregnated by pouring a liquid resin over them. A rubber or plastic blanket is then placed over the fibers, the edges are sealed, and the air in the fibers between the blanket and the mold is driven out by a vacuum pump. Since the rubber blanket is flexible, atmospheric pressure pushes the laminate perpendicularly against the mold. The whole mold is then heated to cure the plastic. The heat and pressure forms a strong laminate in the shape of the mold.

A large number of variations of this process are in use. Cloth or wood veneers may be used in place of mats. Matched and metal molds may be used to apply pressure and heat. Chairs and boxes are made by employing glass fiber mats in matched molds. Rubber bags or diaphragms may be inflated with steam so that they press against the laminate. Glass cloth may be laid over a male mold, impregnated with a plastic solution, and allowed to dry, the process may be repeated until a laminate of the desired thickness is obtained. Different resins which cure to a plastic are used in various modifications depending on temperature available for curing, and the kind of laminating material. Polyesters dissolved in a monomer such as styrene require the addition of a catalyst just before use and can be cured at room temperature. Epoxy resins usually require heating and give an excellent product. Phenolics and melamines require heating.

In the production of standard shapes such as large sheets, cloth or paper is first continuously impregnated with a resin solution by the dip-coating process previously described. The impregnated material is then cut into sheets (usually  $4 \times 8$  ft) and as many as 10–100 of them piled together and placed in a large press where heat and pressure are applied. Paper laminates may have a printed outer layer for decoration. Such sheets are commonly used for table and desk tops. Cloth and fiber laminates are converted into a variety of industrial products by stamping and machining. Tubes and rods are produced by rolling several layers on a mandrel, removing the mandrel, and curing in split molds. Phenolic, urea, melamine, and polyester plastics are used as impregnating resins; curing is done at  $200\text{--}350^\circ\text{F}$ .

Laminates have much higher strength than plastic products which do not contain reinforcing agents. They can be used for building construction, furniture, boats, and other products requiring high strength.

**Extrusion.** An extruder is illustrated in Fig. 3. The die at the point where the plastic emerges may have a variety of shapes in addition to the long slot described under films. Tubes, rods, angles, and many other shapes may be made using special dies. Plastic pipe, tubing, hose, insulated wire, and

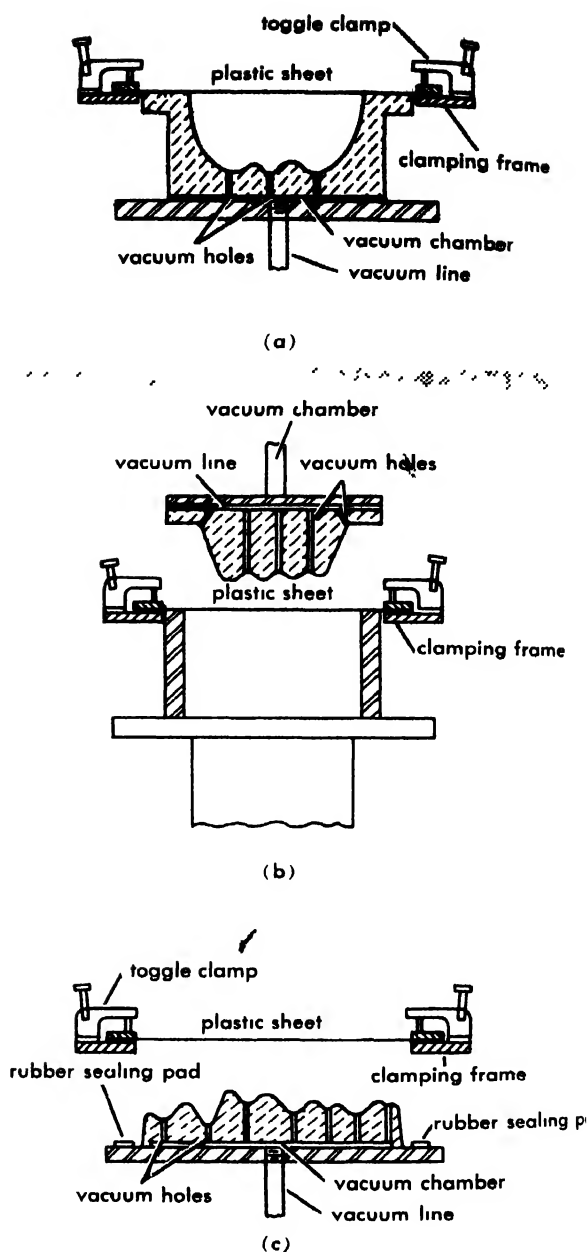


Fig. 8. Three methods of vacuum-forming plastic sheets. (a) Cavity type, (b) Force above sheet, (c) Force below sheet. (Eastman Chemical Products, Inc.)

various profiles for gaskets are made by extrusion. Most thermoplasts can be extruded. Extrusion is relatively inexpensive and the machines are capable of high production rates.

**Sheet-forming.** If thermoplastic sheets are heated, they can be distorted easily and then cooled to retain their shape. In the simplest applications, they may be bent over a mandrel or sealed at the edges and blown against a shaped surface with compressed air. Three methods of vacuum forming are illustrated in Fig. 8. In all variations of vacuum-forming, either a vacuum which exhausts the air behind the sheet through holes in the cavity or a force is used to bring a heated, softened sheet into contact with a surface having

desired profile. Atmospheric pressure forces the softened sheet against the molded surface. Extremely fine detail may be reproduced. The hollow objects made by this method are inexpensive. Special grades of cellulose acetate, vinyl, and polystyrene sheets are made for vacuum-forming. Signs, decorations, boxes, and topographical maps are typical products. See POLYMER. [C.C.WI.]

**Bibliography:** C. C. Winding and R. L. Hasche, *Plastics*, 1947.

### Plate, structural

A flat plate or slab which is supported along the edges, such as the bottom or cover of a tank, cylinder head, bulkhead, or floor panel. A structural plate bends when subjected to forces applied normal to its surface. The bending produces curvature in all planes normal to the plate; this dishing action differs from bending of a beam where curvature occurs in a single plane under symmetrical loads. As for a beam, boundary conditions include various degrees of restraint at the edges, from simple supports to complete fixity against rotation. The analysis of stresses and deflections in a plate, while based on assumptions essentially the same as those used for a beam, is more complicated because of simultaneous bending in all planes. Simplified approximate analyses predict stresses and deflections useful in design. See BEAM.

**Bases for design.** The relative importance of the straining actions involved depends on the thickness and whether the behavior is elastic or inelastic. For thick plates, as for short deep beams,

shearing stresses must be considered, whereas in plates of medium thickness shear stresses can be neglected and only bending action is important. In thin plates, deflection is relatively larger and direct tension due to suspension action is appreciable. For the extreme case of a thin membrane, resistance depends almost entirely on direct tension. Elastic deflection is caused primarily by bending, but during plastic behavior deflections are increased and direct tension resists a progressively larger part of the loads.

The flexure theory for plates, based on assumptions similar to those used for beams, considers the equilibrium of a differential element of a symmetrically loaded plate, included between vertical circumferential and meridional planes, referred to axes of symmetry. The internal stresses and moments are evaluated in terms of angular and linear coordinates of any point of the plate. Application of the general expressions to particular loading and support conditions determines the maximum bending moment. Other methods involve differentiation of an expression for deflection. The stresses  $S$  are calculated by the flexure formula  $S = MC/I$ , where  $M$  is bending moment,  $C$  is the distance from the neutral axis, and  $I$  is the moment of inertia.

The theory assumes elastic action, neglecting shear and direct tension. A simplified approximate approach evaluates the total bending moment on a critical section by statics and the average bending stress by the flexure formula. Maximum stresses are then found by applying correction factors determined by comparison with more exact analysis or experimental results. These analyses assume small deflections and bending as the primary action.

**Circular plate.** A simply edge-supported circular plate can be analyzed approximately by considering the bending of half of the plate about a diametral axis of symmetry.

**Uniform load.** With uniform load, the analysis determines the average bending stress on a diametrical section (Fig. 1). The resultant total bending moment on the diametrical section due to the load and reaction is  $ur^3/3$ . The section modulus is  $rt^3/3$  and the average bending stress at the surfaces is  $S = wr^2/t^2$ , where the symbols are defined as in Fig. 1.

According to the general theory of flexure of plates, the stress reaches a maximum value at the center equal to  $(3/8)(3 + \mu)wr^2/t^2$ , which, for steel with a Poisson's ratio  $\mu = 0.30$ , is 24% greater than the average. Redistribution after local yielding at the center tends to uniformity of stress, and tests have shown that the average stress can be taken as the significant stress.

Elastic deflection at the center is small, except for very thin plates. For a circular plate, simply edge-supported, deflection  $y$  is

$$y = \frac{3}{16}(1 - \mu)(5 + \mu) \frac{wr^4}{Et^3}$$

For  $\mu = 0.30$ ,  $y = (11/16)wr^4/Et^3$  approximately.

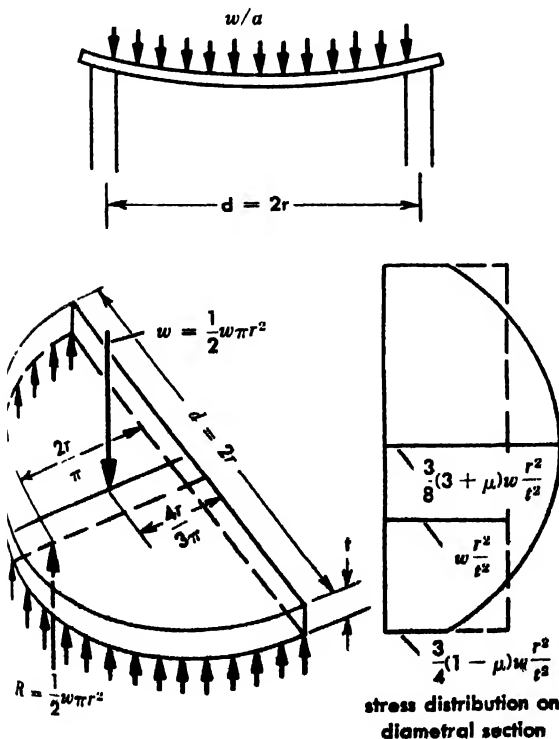


Fig. 1. External forces on the half plate considered in analysis of uniformly loaded circular plate.



**Edges clamped.** Circular plates fixed in direction at the edges and uniformly loaded are analogous to fixed-end beams where negative moments at the ends reduce stress and deflection at the center. The moment is greatest at the edges. The maximum radial bending stress at the edges is  $3wr^2/4t^2$ . For thin plates, the elastic deflection at the center is

$$y = \frac{3}{16} (1 - \mu^2) \frac{wr^4}{Et^3}$$

For  $\mu = 0.30$  this is only 24.5% of the deflection of a simply supported plate.

For thicker plates with  $t/r > 0.1$ , the above value is multiplied by a factor  $C = 1 + 5.72(t/r)^2$ . Complete edge fixity is an ideal condition. Because edge rotation is small for simply supported plates, only a small relaxation of clamping or local yielding will eliminate most of the fixedness, and behavior approaches that of a simply supported plate.

**Central load.** If the central load is distributed over a circular area of radius  $r_0$ , the external forces on a semicircular segment of a simply supported circular plate are the reactions at the rim and the load on the area of application (Fig. 2). The bending moment  $M$  on the diametral section is

$$M = \frac{Pr}{\pi} \left( 1 - \frac{2r_0}{3r} \right)$$

and the average stress is  $S_{avg} = 3P/\pi t^2$ . As the load application area decreases,  $r_0$  approaches zero and the average stress approaches  $3P/\pi t^2$ .

The coefficient to be applied to the average stress to obtain the theoretical maximum stress ( $\mu = 1/3$ ) approaches 1.25 for large values of  $t_0/r$  and is nearly 2.20 for  $r_0/r = 1/10$ .

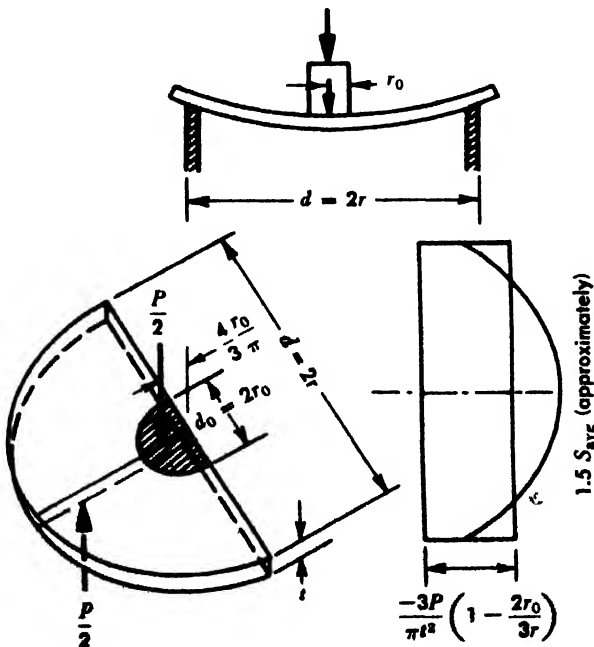


Fig. 2. simply supported circular plate with load concentrated at center.

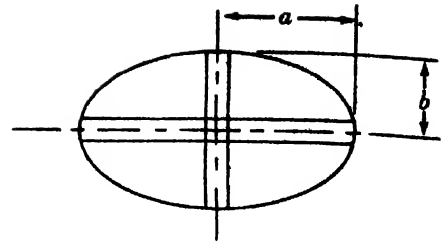


Fig. 3. Elliptical plate.

The theoretical center deflection is

$$y = \frac{3(1 - \mu)(3 + \mu)}{4\pi} \frac{Pr^2}{Et^3}$$

With the edges clamped, the theoretical maximum stress at the center for  $r > 1.7r_0$  is

$$S = -\frac{3(1 + \mu)P}{2\pi t^2} \left( \ln \frac{r}{r_0} + \frac{r_0^2}{4r^2} \right)$$

Yielding at the fixed edges by ductile materials relieves the local stress, and the stresses after redistribution approach those of the simply supported plate.

When  $r_0/r$  is small, the center deflection is

$$y = \frac{3(1 - \mu^2)Pr^2}{4\pi Et^3}$$

This deflection increases with yielding at supports.

**Center support.** A circular plate, supported at its center, with uniform load has a maximum theoretical stress at the center of

$$S_{max} = \frac{3wr^2}{2t} \left[ (1 + \mu) \ln \frac{r}{r_0} + \frac{1}{4} (1 - \mu) \left( 1 - \frac{r_0^2}{r^2} \right) \right]$$

For  $r_0/r$  small, the term  $r_0^2/2$  is negligible. Under this condition, the center deflection is

$$y_{max} = \frac{3}{16} (1 - \mu)(7 + 3\mu) \frac{wr^4}{Et^3}$$

When  $\mu = 1/3$  this reduces to  $y_{max} = wr^4/Et^3$  which is three-fifths of the deflection for the same central load on an edge supported plate.

**Elliptical plate.** In an elliptical plate simply supported at its edges and with uniform load, the maximum bending stress and curvature occur in the direction of the minor axis (Fig. 3). If the ellipse is elongated with  $a$  very much greater than  $b$  the plate action approaches that of a simply supported beam with a span of  $2b$ . A central strip of unit width along the minor axis resists a maximum moment of  $w b^2/2$  and the maximum stress in the plate is  $3w b^2/t^2$ . If  $a$  equals  $b$ , the plate is circular and the maximum average stress is  $w b^2/t^2$ . For intermediate dimensions the coefficient of  $w b^2/t^2$  varies between 1 and 3. An approximate expression for maximum stress, in the direction of the minor axis, is

$$S_{max} = \frac{3a - 2b}{a} \frac{w b^2}{t^2}$$

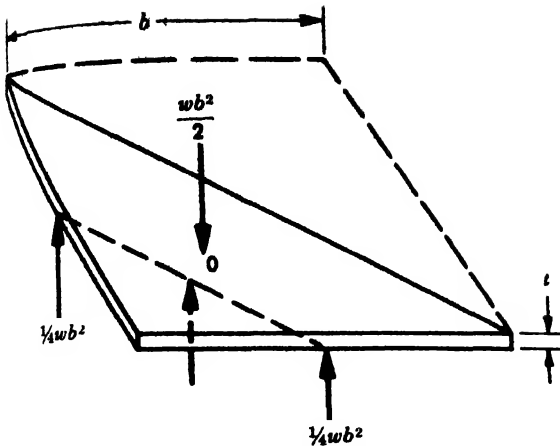


Fig. 4 Forces on triangular half plate serve to analyze uniformly loaded square plate.

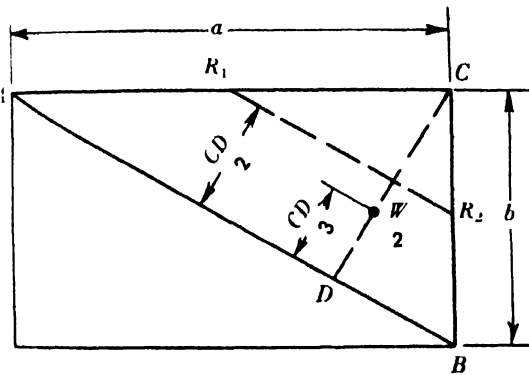


Fig. 5 Rectangular plate simply supported and uniformly loaded

**Square plate.** When simply supported at all edges with uniformly distributed load, a square plate can be analyzed as follows. The critical section is taken along the diagonal of the square, about which maximum stress occurs. The bending moment on a diagonal section is found by considering the forces on the triangular half plate (Fig. 4). The resultants of the applied load and reactions and their locations determine the bending moment on the diagonal section which is  $M = wb^3/12\sqrt{2}$ . The average stress on the diagonal section is  $wb/4t^2$  which is the same as for a circular plate with diameter equal to the side of the square.

With fixed edge supports, the average stress on the diagonal section is taken as  $wb^2/5t^2$ .

**Rectangular plate.** A simply supported rectangular plate with uniformly distributed load again has the critical section along the diagonal. Because of symmetry, the resultant edge reactions are at the centers of the sides (Fig. 5). On the half rectangle applying under static equilibrium, the average stress across the diagonal section is

$$S_{avg} = \frac{1}{2} \frac{a^3}{a^2 + b^2} \frac{wb^2}{t^2}$$

For  $a = b$ , the expression reduces to that for a

square plate. For a rectangle, the stress is always greater in the direction of the shorter span. For  $a/b$  very large, the average stress across a midsection parallel to the sides approaches that of a simply supported plate with span =  $b$  and  $S_{avg} = 3wb^2/4t^2$ .

**Concentrated load at center.** A simply supported square plate with a concentrated load  $P$  applied to a central area of diameter  $d_0$  has an average bending stress on the diagonal section  $S = 3P/4t^2$ . However, the maximum stress at the center found from the theory of flexure of plates by H. M. Westergaard is considerably larger,  $S = 2.64 P/t^2$  for  $\mu = 1/4$ . The higher stresses are localized and plastic yielding, which causes redistribution, does not greatly affect the plate as a whole. For brittle materials or cyclic loading involving fatigue, the high localized stress is important.

**Continuous plates.** In some cases the plate extends beyond its supports. Important examples are flat floor slabs supported by equally spaced columns and "stayed" flat steel plates used as boiler heads.

For plates supported by rods or stays attached so as to divide the plate into equal square panels, the maximum bending stress, according to Unwin, is

$$S = \frac{2}{9} w \frac{a^2}{t^2}$$

where  $t$  is thickness,  $a$  is distance between supports (side of the panel) and  $w$  is uniformly distributed load per unit area. All units are in inches.

A floor slab may be supported by circular columns  $L$  distance apart and dividing the continuous plate into square panels (Fig. 6). The analysis involves moments at five edges of a quarter panel. The distribution of moments along the exterior edges is unknown and the problem is statically indeterminate. From the theory of flexure, Westergaard found expressions for the moment per unit width at various locations in the panel in terms of the average moment per unit width along one en-

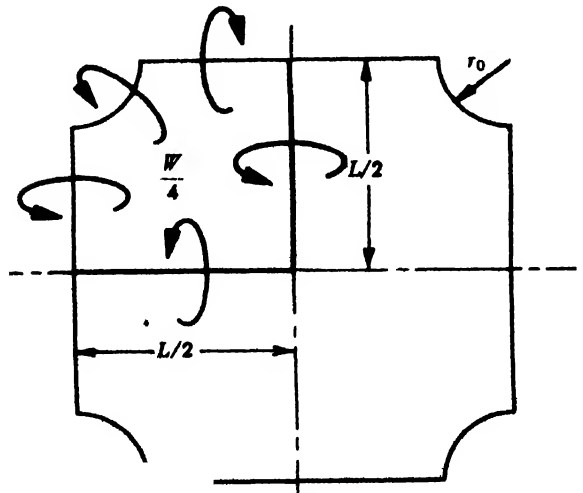


Fig. 6. Continuous plate.



result of erosion of ultrabasic rocks and are found in Alaska, Colombia, and the Soviet Union. In Canada, the platinum metals occur in the nickel-copper sulfide ores, and the production is dependent entirely upon the demand for nickel. The platinum content of these ores varies somewhat, but is of the order of 1 part per million, practically all of which is recovered. The platiniferous ores of South Africa contain minor amounts of copper and nickel, but these ores are mined primarily for their platinum content. South African gold mines yield small amounts of osmiridium, containing about 35% osmium and 30% iridium, and iridosmine, which contains about 40% iridium. Small amounts of the platinum metals, palladium in particular, are recovered during the electrolytic refining of copper. The major sources of platinum metals are Canada, South Africa, and the Soviet Union.

The principal platinum minerals are platinum arsenide (sperrylite), platinum sulfide (copperite), platinum palladium nickel sulfide (braggite), native platinum, osmiridium, and iridosmine.

**Uses.** The platinum group metals have wide industrial use because of their catalytic activity, chemical inertness, and high melting points. As a catalyst platinum is used in hydrogenation, dehydrogenation, isomerization, cyclization, dehydration, dehalogenation, and oxidation reactions. Finely divided platinum on aluminum oxide is used to upgrade the octane rating of gasoline by catalytically accelerating a complex series of dehydrogenation, hydrogenation, isomerization, and cyclization reactions. Platinum-rhodium alloy gauzes are used in the catalytic oxidation of ammonia to form nitric acid or oxides of nitrogen. When this process is carried out in the presence of methane, hydrocyanic acid is formed. Platinum is a catalyst in the older contact process for making sulfuric acid and the conversion of carbon monoxide to carbon dioxide by combination with oxygen, the recombination of hydrogen and oxygen to form water, the reduction of nitro groups, and the removal of nitrogen(II) oxide from gas streams by reaction with hydrogen to form nitrogen and water. Platinum and platinum alloys are used in many special applications including insoluble anodes, jewelry, spinnerets used in synthetic fiber extrusion, equipment for melting, stirring, and extruding molten glass, thermocouples, resistance thermometers, electrical contacts, dental and medical devices, corrosion-resistant laboratory utensils, electric-furnace windings, catalytic gas ignitors, pressure rupture disks, and grids of special purpose vacuum tubes.

**Alloys.** Pure platinum is soft and ductile. Moderate hardness is obtained with the addition of 0.5% iridium or 3.5% rhodium. Such alloys are used in the manufacture of laboratory utensils. For jewelry, a 5–10% iridium or a 5% ruthenium alloy is used. The 10% rhodium alloy is used in thermocouples, ammonia-oxidation catalysts, spinnerets, furnace windings, and for molten glass handling equipment. A 13% rhodium alloy is also used for thermocouples. A 4% tungsten alloy is used as

heavy-duty electrodes in aircraft and industrial spark plugs.

**Chemical and physical properties.** Platinum has a density of 21.45 g/cm<sup>3</sup> at 20°C, melts at 1769°C, and boils at 4530°C. Its electrical resistivity is 10.6 microhms/cm at 0°C. Radioisotopes of the following mass numbers are known: 187, 188, 189, 190, 191, 193, 197, and 199. The stable isotopes of platinum have the mass numbers 192, 194, 195, 196, and 198. The natural abundances of these are 0.8, 30.2, 35.3, 26.6, and 7.2% respectively. Platinum is not attacked at room temperature by any single acid but it is readily dissolved by hot aqua regia. Although it is resistant to hydrogen chloride at elevated temperatures, it does react with chlorine at about 500°C. It is resistant to mercury, fused sulfates, chlorides, and carbonates. Platinum exhibits valence states of 1+, 2+, 3+, 4+, and 6+. The 2+ and 4+ valences are the most common. Platinum can be made into a spongy form by thermally decomposing ammonium chloroplatinate or by reducing it from an aqueous solution. In this form, it exhibits a high absorptive power for gases, especially oxygen, hydrogen, and carbon monoxide. The high catalytic activity of platinum is related directly to this property. Hydrogen diffuses through heated platinum. Platinum strongly tends to form coordination compounds.

**Metallurgical extraction.** There is no routine method for the extraction of platinum. The method used is dependent upon the starting material, which may be scrap or used catalyst, slimes resulting from nickel or copper processing, crude platinum, or the very refractory osmiridium or iridosmine. The platinum can be extracted with aqua regia in some cases. In other cases, it may be necessary to fuse the ore with a suitable flux and to collect the platinum group metals in a carrier such as copper or lead. When a hydrochloric acid solution of platinum is oxidized and then made basic, most of the impurities precipitate as hydroxides, and the platinum remains in solution. A similar purification can be obtained by converting the platinum to the hexanitrito complex which does not precipitate in basic solution. When a hydrochloric acid solution of platinum(IV) is treated with ammonium chloride, the platinum is almost completely precipitated as ammonium chloroplatinate. The chloroplatinate is readily converted to the metal by thermal decomposition. Platinum can also be reduced to the metal from aqueous solutions of its salts by zinc, magnesium, iron, or aluminum. The platinum in dilute aqueous solutions can also be precipitated as the sulfide and thereby concentrated. The refining of a specific material consists of a combination of these methods. In addition to separating the precious metals from all other impurities, it is necessary to separate them from each other. The refining of precious metals requires a great deal of flexibility on the part of the refiner. It is not unusual for a portion of the material to be recycled in the refinery because of the lack of suitable reactions for quantitative separations.

**Principal compounds.** Platinum dioxide,  $\text{PtO}_2$ , is a dark-brown, insoluble compound, commonly known as Adams catalyst. It is prepared by fusing chloroplatinic acid with sodium nitrate at  $500^\circ\text{C}$ . Solution of the melt in water separates the salts from the insoluble platinum dioxide. Platinum(II) chloride,  $\text{PtCl}_2$ , is an olive-green, water-insoluble solid. It is made by heating platinum in chlorine at  $500^\circ\text{C}$ , or by the thermal decomposition of chloroplatinic acid. Platinum(II) chloride dissolves in hydrochloric acid to form chloroplatinous acid,  $\text{H}_2\text{PtCl}_4$ , which cannot be isolated but which forms soluble salts such as potassium platinum(II) chloride,  $\text{K}_2\text{PtCl}_6$ . Chloroplatinic acid,  $\text{H}_2\text{PtCl}_6$ , the most important platinum compound, is made by dissolving platinum in aqua regia, destroying the nitric acid by evaporation from a hydrochloric acid solution, and evaporating the solution. The acid is isolated as a hydrate,  $\text{H}_2\text{PtCl}_6 \cdot 6\text{H}_2\text{O}$ . The red-brown crystals are very soluble in water. Ammonium chloroplatinate,  $(\text{NH}_4)_2\text{PtCl}_6$ , is a lemon yellow, crystalline, relatively insoluble solid made by adding ammonium chloride to a solution of chloroplatinic acid. Compounds such as dichlorodiamino platinum,  $\text{Pt}(\text{NH}_3)_2\text{Cl}_2$ , exhibit cis-trans isomerism because of their planar configuration.

**Analytical techniques.** If platinum is to be determined in an ore-type residue, it is collected in lead or silver while fusing the material with a suitable flux. Recent data indicates that iron, nickel, or copper may be superior collection media. The metallic residue is dissolved in aqua regia, and the nitric acid is destroyed by evaporation from hydrochloric acid. The platinum is determined by reading the optical density of its stannous chloride complex. If it is necessary to remove impurities from the platinum, this may be accomplished by passing a solution of the anionic hexachloroplatinum complex through a cation exchanger. The impurities are generally cationic and are therefore removed by the resin. Hydrolysis or nitritation may also be used. Potassium iodide or thiosemicarbazide may be used for the colorimetric determination of platinum. Thiophenol has been used for its gravimetric determination. See HYDROGENATION; IRIIDIUM; OSMIUM; PALLADIUM; RHODIUM; RUTHENIUM. [F. A. HAMMOND]

**Bibliography:** A. D. Lumb, *The Platinum Metals*, 1920.

## Platyasterida

An order of Asteroidea in which traces of metapinnules persist, the ossicles of the arm skeleton being arranged in two growth gradient systems. One system causes the ossicles to lie in transverse rows, like the corresponding structure of somasteroids, and the other causes the ossicles simultaneously to lie in longitudinal rows, as in all other asterooids. See ASTEROIDEA; SOMASTEROIDEA.

Fossil members of the order are known from the Ordovician onward, but recently the extant family of starfishes Luidiidae has been recognized as living representatives of the Platyasterida. The fossils first studied happened to be severely compressed by

geological processes, thus leading to the (incorrect) inference that no ambulacral groove was developed; in fact, both the fossil forms and their living representatives have a well-developed ambulacral groove, as in other starfishes. One row of marginal plates is developed around the arm. See PAXILLOSINA; PHANEROZONIDA. [H. B. FELL]

**Bibliography:** H. B. Fell, Phylogeny of sea stars, *Phil. Trans. Roy. Soc. London, Ser. B*, 246:381-485, 1963.

## Platyropa

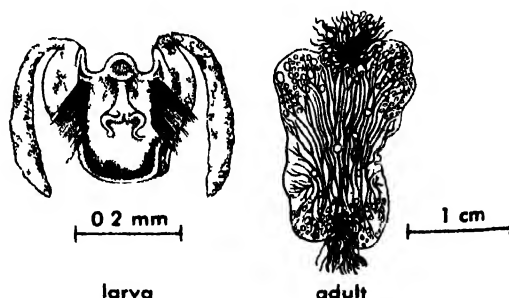
A suborder of the ostracods of the class Crustacea. This suborder contains the single family Cytherellidae. Heart and eyes are lacking. The exopodite and endopodite of the second antennae are well developed, having broad, flattened podomeres, or appendage segments, that distinguish members of this suborder from all other ostracods. Both pairs of antennae can be extended anteriorly but are not used for locomotion. There are three pairs of postoral appendages, none of them leglike. In the female the last pair are rudimentary.

Some species have been found only in the profundal sediments of the Atlantic Ocean and the Mediterranean Sea; other species inhabit the Caribbean Sea, the Arabian Sea, and the Gulf of Mexico. The genus *Cytherella*, created by J. Bosquet in 1852, includes present-day forms but was established to accommodate fossil forms taken from the Tertiary deposits of France and Belgium. See OSTRACODA. [E. FERGUSON]

**Bibliography:** W. A. Tressler, Marine Ostracoda from the Gulf of Mexico, *U.S. Fish Wildlife Serv. Fishery Bull.*, 55:429-437, 1954.

## Platyctenea

An order of the ctenophores whose members are sedentary or parasitic. They often lack ribs in the adult stage. They are flattened due to the shorten-



*Coeloplana bocki*.

ing of the main axis and by the extension of the pharynx into a creeping sole. These organisms are beautifully colored. The primary tentacles are well-developed and the canals branch profusely to form a network. Some are viviparous with the cydippid embryos developing in brood chambers formed by the expansion of the canals. *Ctenoplana*, *Coeloplana*, *Tjalfiella*, *Lyrocteis*, and *Gastroides* are representative genera. See CTENOPHORA; TENTACULATA. [T. KOHAI]

## Platyhelminthes

A phylum of the invertebrates, commonly called the flatworms. They are bilaterally symmetrical, non-segmented, dorsoventrally flattened worms characterized by lack of coelom, anus, circulatory and respiratory systems, and exo- or endoskeleton. They possess a protonephridial excretory system, a complicated hermaphroditic reproductive system, and a solid mesenchyme which fills the interior of the body. Three classes occur in the phylum: (1) the Turbellaria, mainly free-living, predaceous worms; (2) the Trematoda, or flukes, holozoic ecto- or endoparasites; and (3) the Cestoda, or tapeworms, saprozoic endoparasites in the enteron of vertebrates, whose larvae are found in the tissues of invertebrates or vertebrates.

**Morphology.** Flatworm tissues and organs are derived from three germ layers. Most flatworms have adhesive or attachment devices. In turbellarians these are glandular or muscular; in flukes and cestodes they are suckers and cuticularized hooks or spines. Muscular tissue occurs in mesenchymal layers and permits rapid change in body form. Numerous receptor cells (tangoreceptors and hemoreceptors) and sense organs (tentacles, statocysts for balance, and ocelli for photoreception) occur on the head and body. Many of these are reduced or absent in parasitic forms. Primitive flatworms possess a coelenterate-like nerve net, whereas higher forms have a brain in a head region as well as two main longitudinal nerve cords which resemble a ladder because of the cross commissures. A digestive system is present in turbellarians and trematodes, but is lacking in cestodes. A mouth serving also for egestion, is situated either anteriorly or ventrally. It is often provided with an aspirating sucker or pharynx by which small organisms, juices of larger ones, or host tissues are ingested. The gut is sac-like or branched and lined with a single layer of cells often packed with granules. Digestion is inter- or intracellular. The protonephridial excretory system is of uniform construction in the phylum. It consists of ramifying, blind tubes capped with large cells, called flame cells, each bearing a tuft of cilia which projects into the lumen. The tubes course through the mesenchyme and discharge at the surface by means of one or more openings. The function of this system is not well understood. The reproductive system reaches morphological complexity beyond that found in other phyla. In each reproductive unit, which is the entire organism in Turbellaria and Trematoda and is the proglottid in the Eucestoda, one to several testes and ovaries occur as well as various accessory reproductive organs.

**Reproduction.** The platyhelminthes reproduce both sexually and asexually. In sexual reproduction, fertilization is internal, following copulation or hypodermic insemination. Cross or self fertilization may occur. Except in acoel and polyclad flatworms, the eggs are ectolecithal, that is, the ova are invested with yolk cells and the egg mass is enclosed in a capsule. In turbellarians, capsules are

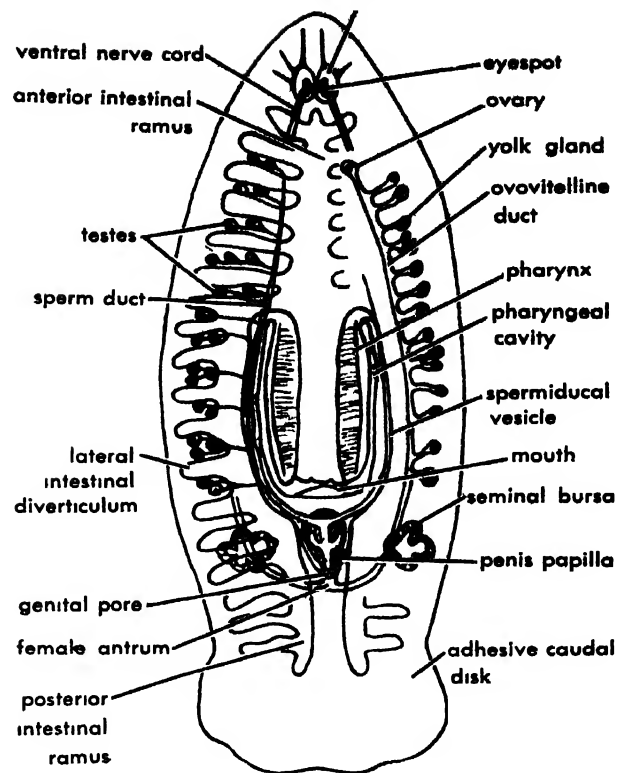


Fig 1. *Bdelloura candida* (Tricladida), ectocommensal on the kingcrab, *Limulus*. Complete digestive and male systems shown on left, female systems on right.

often deposited in cocoons in which the juveniles develop. Eggs of trematodes and cestodes emerge from the hosts and are either eaten or hatch in water as free swimming larvae. They then either actively penetrate or are passively ingested by intermediate hosts.

Asexual reproduction is of frequent occurrence in the phylum. Many turbellarians reproduce by fragmentation or binary fission; in some, chains of individuals are temporarily produced. Trematodes reproduce asexually (polyembryony) in their larval stages, as do some tapeworms such as *Echinococcus* and *Multiceps*. Formation of proglottids is an asexual process, resulting from activity of a proliferating zone in a neck region.

**Economic and biological importance.** Turbellaria are widespread in fresh water and the littoral zones of the sea, while one group of triclads occurs on land in moist habitats. Turbellaria have been used to study regeneration, including the effects of chemicals and radiation upon the process. They have also been used in axial gradient research. Adult trematodes occur on, or in, practically all tissues and cavities of the vertebrates on which they feed. They are responsible for troublesome diseases in man and animals. Larval flukes are frequent in mollusks, mainly gastropods, and occasionally occur in pelecypods. Vector hosts, such as insects and fish, are often interpolated between mollusk and vertebrate. Adult tapeworms, living in the enteron or the biliary ducts, compete with the host for food



and accessory food factors such as vitamins. Larval tapeworms reside chiefly in arthropods, but larvae of one group, the Cyclophyllidae, develop in mammals, which may be severely impaired, or even killed, by the infection. Investigations utilizing flukes and tapeworms have given clearer insight into the host-parasite relationship.

**Classification.** The following classification is based on that of L. Hyman, with modifications. Separate articles appear on each group.

**Phylum Platyhelminthes**

**Class Turbellaria**

- Order Acoela
- Order Rhabdocoela
- Order Alloeocoela
- Order Tricladida
- Order Polycladida

**Class Trematoda**

- Order Monogenea (Heterocotylea)
- Order Aspidogastrea (Aspidobothria; Aspidocotylea)
- Order Digenea (Malacocotylea)

**Class Cestoda**

**Subclass Cestodaria**

- Order Amphilinidea
- Order Gyrocotylidea

**Subclass Eucestoda**

- Order Tetraphyllidea (Phyllobothrioidea)
- Order Lecanicephaloidea
- Order Proteocephaloidea
- Order Diphyllidea
- Order Trypanorhyncha (Tetrarhynchoidea)
- Order Pseudophyllidea (Bothriocephaloidea)
- Order Nippotaeniidea
- Order Cyclophyllidea (Taenioidea)
- Order Aporidae

**Turbellaria.** This class, considered ancestral to the other two, is the most primitive of the phylum: five orders are recognized. The body is cylindrical or flattened and may be less than 1 mm to over 50 cm long. The members are chiefly free-living, but commensal and parasitic species occur. A ciliated epidermis may uniformly cover the body or may be restricted to the ventral or other surfaces. Peculiar rodlike rhabdoids of uncertain function are embedded in the epidermis. Characteristic color patterns are imparted to some species by chromatophores, while other species contain symbiotic algae and are green or brown. The order Acoela contains small marine forms lacking an enteron. They digest food within phagocytic mesenchyme cells and lack an excretory system; statocysts are usually present. The order Rhabdocoela consists of small marine or fresh-water forms with a simple blind intestine. One suborder, the Temnocephalida, occurs as ectocommensals on fresh-water animals. The order Alloeocoela comprises small marine or fresh-water organisms with saclike or lobulated gut, while the order Tricladida consists of larger forms, either marine, fresh-water, or terrestrial, which have a three-branched intestine (Fig. 1). The order Polycladida contains the marine forms of elongate or leaflike shape with few to

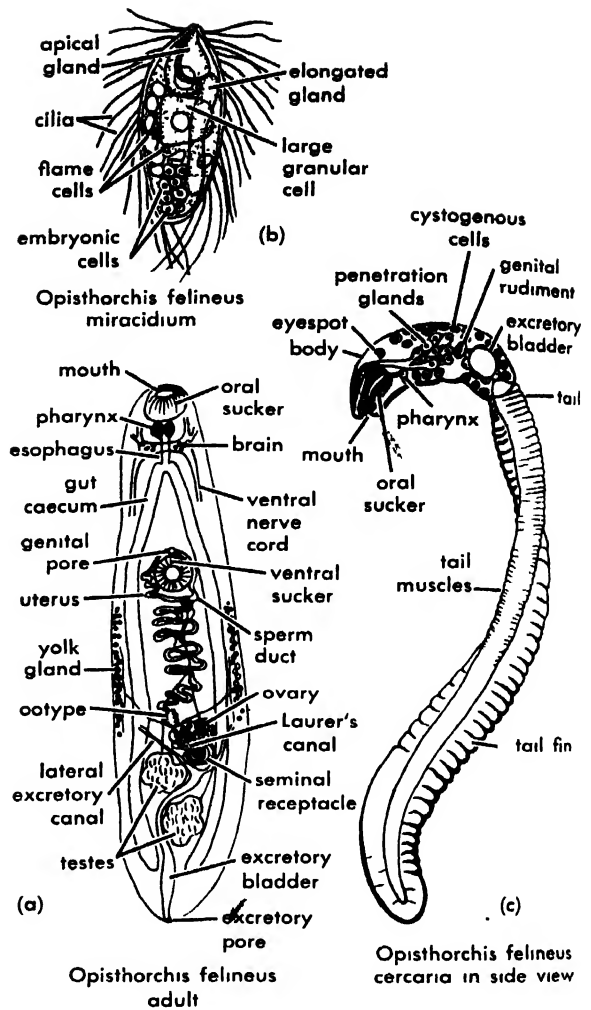


Fig. 2. Stages in life cycle of *Opisthorchis felinus* (a) Adult. (b) Miracidium. (c) Cercaria. (Redrawn from H. Vogel, *Zoologica*, 33:1-103, 1934)

many eyes near the tentacles and on the body margin. Many intestinal rami branch from a central digestive cavity above the pharynx.

**Trematoda.** This large class consists of parasites covered with a secreted cuticula. Three subclasses or orders occur here: (1) Monogenea, found mainly as ectoparasites on cold-blooded vertebrates, have direct and simple life cycles; (2) Aspidobothria endoparasites of invertebrates and vertebrates, are characterized by their alveolated ventral attachment surface and by their either direct or complicated life cycle; (3) Digenea, endoparasites of vertebrates, have two or more hosts involved in their life cycle (Fig. 2). The intermediate host is a mollusk.

**Cestoda.** These highly specialized animals are parasites of most vertebrate classes. The typical tapeworm, subclass Eucestoda, consists of an attachment organ, the scolex, and a series of similar reproductive segments, the proglottids. Members of the other subclass, Cestodaria, are nonsegmented and occur in the intestine and coelom of fishes; the larvae occur in crustaceans.

In Eucestoda the ripe eggs, or hatched ciliated larvae (coracidia), are ingested by intermediate hosts in which six-hooked embryos, the onco-spheres, transform into larval forms such as pro-cercoids, plerocercoids, cysticeroids, cysticeri, and other specialized types. Final hosts become infested by ingesting these larvae. One species, *Hymenolepis nana*, may have a direct life cycle.   
 SEE ACCELOMATA. [C. C. CRONEIS]

## Playa

The low, essentially flat part of a basin or other undrained area in an arid region. In heavy rains the playa may be temporarily covered with a shallow sheet of water, and is then a playa lake.

Five principal types of playa have been recognized by Richard O. Stone. The dry playa develops where the water table is below capillary reach of surface; its surface is hard, flat, and smooth, composed of silt and clay. The moist playa or salina occurs where the water table is within capillary reach of surface; this type may be further subdivided into (1) salt encrusted, where the water



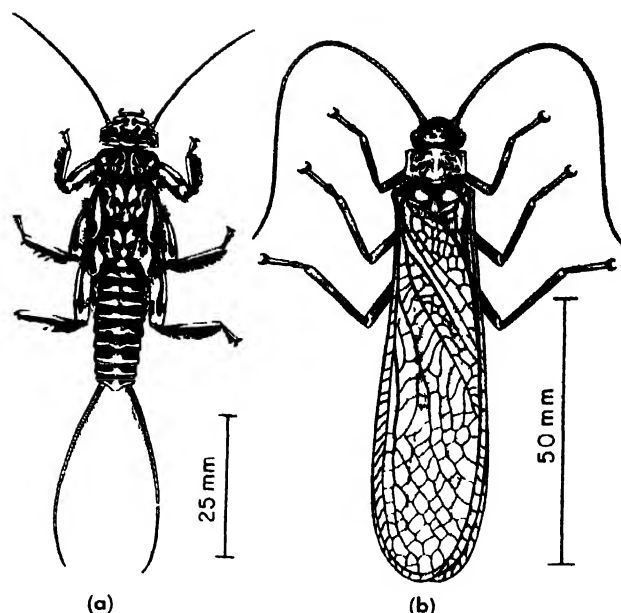
Light colored saline deposits encrust the lowest parts of the basin or bolson filling in Death Valley, Califor-

table is at the surface or so near it that salt water evaporating on the surface leaves a salt crust; and (2) clay encrusted, where the water table is near the limit of capillary action and salt brought to the surface is mixed with silt and clay, forming a puffy surface ("self-rising ground"). The crystal body playa is essentially a massive body of crystalline salt at or very close to the surface. See EVAPORITE (SALINE). The compound playa results from a water table at different levels in different parts; these exhibit characteristics of dry playa in one part and moist playa in another. The lime-pan playa is formed in basins receiving drainage from limestone terrain and has a floor of hard travertine.   
 SEE DESERT EROSION FEATURES. [T. CLEMENTS]

**Bibliography:** T. Clements et al., *A Study of Desert Surface Conditions*, U.S. Army Tech. Rept. EP-53, 1957.

## Plecoptera

An order of insects known as the stone flies. They are among the most primitive of insects. Except for wings and tracheal gills, there are relatively slight differences between aquatic immature species and aerial mature ones. Striking differences between immature stages are characteristic of specialized insects, such as butterflies. In stone flies, the soft



Plecoptera. (a) Nymph of *Perla* showing gills at bases of legs. (b) Adult of *Pteronarcys*, one of the largest stone flies (A. H. Morgan, *Field Book of Ponds and Streams*, Putnam, 1930)

flattened body, strong walking legs, paired claws on each foot, biting mouthparts, rusty blacks, dull yellows, and browns are characteristic of both phases of life.

Stone flies live in clean swift streams through their immature, nymphal life which extends from 1 to 3 years. A complete life history of one species has been observed in which the nymphs went through 22 instars, the form assumed by an insect during developmental stages. The aquatic life of stone flies ends when the mature nymphs climb onto rocks or plants, and shed their skins. Adults are poor fliers and are often found creeping about on stream banks and shrubbery. The adult life is relatively short. Mating occurs near the ground, never in flight. Stone flies that inhabit large streams deposit their eggs directly into swirling water in clutches of perhaps 1400, while those of small streams place the eggs at the water's edge.

Filamentous tracheal gills are characteristic of the nymphs except those that live in highly aerated water and breathe through the body wall. In various species, gills are attached at the neck, on the sides of the thorax, and on the sides and tip of the abdomen. In several species, such nymphal structures persist in the adults, but do not function.

Adult stone flies hold their wings close to their backs when at rest or walking. The hindwings are pleated and hidden. The name Plecoptera, meaning pleated wings, comes from this habit. Folded wings constitute a step in evolution never achieved by mayflies, or even by butterflies.

In Illinois, stone flies emerge from the water in every month of the year. Successions of species reach their peaks of abundance from November to March. Nymphs and adults of several species are

plant-feeders; others are carnivorous. Their diet can be tested only by examining the stomach content of newly collected specimens. See INSECTA.

[A. H. MORGAN]

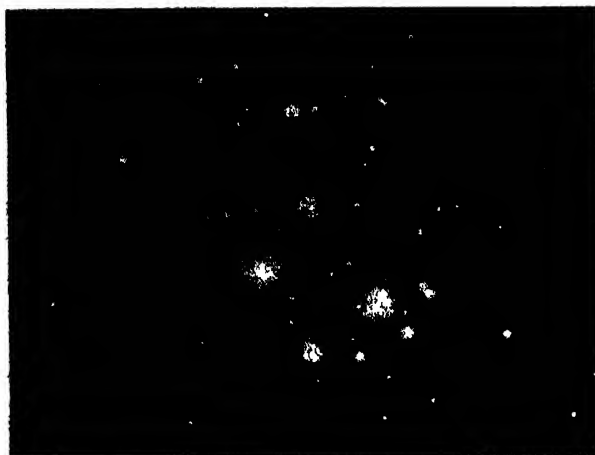
## Plectoidea

A superfamily of small free-living nematodes with mushroom-shaped amphids and reflexed ovaries. Members of the first of the two families, Plectidae, are mainly found in soil and fresh water, whereas those of the second, Camacolaimidae, are mainly marine. Some of the fresh-water species can withstand lengthy desiccation, and become active again when moistened. Food habits are unknown, but plectoids are probably microbivorous. See NEMATODA.

[H. E. WELCH]

## Pleiades

A beautiful group of stars resembling a little dipper, in the constellation of Taurus, known since earliest records. The Pleiades is a typical galactic cluster; it contains several hundred stars within a radius of  $1^\circ$  from Alcyone. Its distance is 410 light years, its linear diameter about 15 light years. The brightest stars are blue, of B type. The cluster is



The Pleiades. (Lick Observatory photograph)

permeated with diffuse nebulosity. Though early accounts refer to the Pleiades in terms of seven stars, only six are now conspicuous to the unaided eye, which raises a theory that one, the lost Pleiad, has faded.

[H. S. HOGG]

## Pleiotropism

A gene is said to be pleiotropic if it has more than one phenotypic effect. Manifold gene effects are often striking; less obvious ones can be detected in most, if not in all, genes by appropriate methods of observation. Pleiotropic gene effects often bear an obvious relationship to each other. They are said to be coordinated if they are brought about by the same mechanism in different parts of the body. For instance, in *Drosophila*, many genes which affect eye color have a similar effect on testis coloration. In other cases, pleiotropic gene effects are subordinated to each other, one effect being the cause of another, a relationship called a hierarchy of

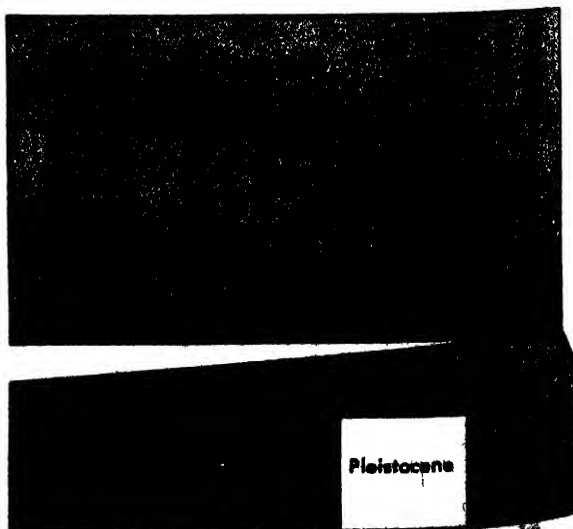
causes. For instance, imperfect plumage in the Frizzle fowl is the cause of numerous physiological consequences, all of which are traceable to increased loss of body heat. Where a pleiotropic pattern can be reduced to either coordinated or subordinated gene effects or to both, unity of primary gene action may be inferred, and the pleiotropism may be regarded as spurious. Genuine pleiotropism would involve the existence of more than one primary gene action. No examples of genuine pleiotropism have so far been demonstrated though, of course, in many pleiotropic systems the causal relationships of the effects are not yet understood. See GENE ACTION.

[H. GRÜNEBFRG]

## Pleistocene

A term referring to a sequence of geologic deposits (the Pleistocene Series) and also the time (the Pleistocene Epoch of the Quaternary Period) during which those deposits were made. Introduced by the British geologist Sir Charles Lyell in 1839 it has been defined on various bases, such as its content of fossil mollusks, its fossil mammals, and its evidence of glacial climate. No single definition is universal. In most European countries and by many authorities in the United States, the Pleistocene is considered to begin with the first appearance of the horse, cattle, and elephant in the specific sense. According to the U.S. Geological Survey, the Pleistocene includes the Great Ice Age, or Glacial Epoch, and possibly some preglacial time. In sedimentary cores raised from the deep sea floor, the Pleistocene Series is identified on a basis of inferred water temperatures. Although Pleistocene time has been regarded by some as including the present, more common usage considers it as ending with the end of the Ice Age, after which Recent, or Holocene, time began. See GLACIAL EPOCH; QUATERNARY; see also MARINE SEDIMENTS.

The Pleistocene deposits include a very large variety of sediments. In middle and high latitudes glacial deposits are prominent among them. In most places these are unconsolidated or semicon-



solidated, and blanket the underlying bedrock over wide areas. The length of Pleistocene time has not yet been measured but is generally estimated to be of the order of 1,000,000 years. [R. F. FLINT]

### Pleochroic halos

Small halos of color or color differences that are sometimes observed around inclusions in minerals. They were first noted by Harry Rosenbusch (1873) around cordierite and were later reported by many observers in a great many minerals. If halos occur in doubly refracting substances, they may show pleochroic discoloration, and the term pleochroic halos is loosely applied to such small colored halos generally. See CORDIERITE; PLEOCHROISM.

**Distribution and description.** The halos are found only around minute inclusions of certain minerals, especially zircon, allanite, monazite, and others known to contain minor amounts of uranium or thorium. Halos have been reported in a great many rock-making minerals, including amphiboles, pyroxenes, and micas. See RADIOACTIVE MINERALS.

The halos are usually spheroidal (circular, as seen in microscopic section), sometimes consisting of several concentric rings, and are fairly sharply bounded. The outermost ring in biotite does not attain a diameter of more than 0.04 mm.

**Origin and interpretation.** J. Joly in 1907 was the first to ascribe pleochroic halos to the effect of irradiation with  $\alpha$ -particles. He and others later attempted with only limited success to explain the details of the ring structure by the ranges (length of tracks) of  $\alpha$ -particles from the several sources in the uranium or thorium series in the mineral affected. Changes in refringence or birefringence (either increase or decrease) may be associated with the coloring, and it has been suggested that the formation of pleochroic halos is comparable to the radiation damage in minerals called metamictization. However, pleochroic halos are most widespread in minerals such as biotite, never known in the metamict state. The phenomenon is doubtless more closely related to other types of coloration induced by radiation. Pleochroic halos have been produced artificially in various materials by E. Rutherford and others. See BIREFRINGENCE; GEM, MANUFACTURED; METAMICT STATE.

It has been suggested that the halos might offer a means for estimating the ages of minerals. This seems unlikely, since there is a limit to the coloration that can be produced in any material; moreover, there is some indication that reversal can occur on prolonged exposure and that the color can be dissipated by heat. [A. P. PABST]

**Bibliography:** D. E. Kerr-Lawson, Pleochroic haloes in biotite, *Univ. Toronto Studies, Geol. Ser.*, 27:15-27, 1928; J. Orsel, L'état métamict, *Bull. soc. belge géol. paleontol. et hydrol.*, 65:165-194, 1956; P. Ramdohr, Neue Beobachtungen an radioaktiven Höfen in verschiedenen Mineralien mit kritischen Bemerkungen zur Auswertung der Höfe zur Altersbestimmung, *Geol. Rundschau*, 49:253-263, 1950; K. Rankama, *Isotope Geology*, 1950.

### Pleochroism

In some colored transparent crystals the color is quite different in different directions through the crystals. This effect is sometimes called pleochroism or trichroism. In such a crystal the absorption of light is different for different polarization directions. Tourmaline offers one of the best known examples of this phenomenon. In colored transparent tourmaline the effect may be so strong that one polarized component of a light beam is wholly absorbed, and the crystal can be used as a polarizer.

For a fuller discussion of the effect, see DI-CHROISM; TRICHROISM. [B. H. BILLINGS]

### Pleuracanthodii

An order of Paleozoic sharklike fishes abundant in fresh-water deposits of the Carboniferous and early Permian. Although primitive in cranial



*Pleuracanthus*, a Carboniferous and Permian sharklike form; perhaps 2½ ft long. (After Fritsch)

structures, the pleuracanthos differ notably from other sharks in several regards. The teeth are 2-pronged; there is a long spine projecting upward and backward from the posterior brain case; the tail extends directly backward, in contrast to the upturned heterocercal caudal fin of other sharks; and the paired fins have the archipterygial skeletal pattern of a central axis and side branches, in contrast with a fan-shaped arrangement in typical sharks. See ELASMOBRANCHII FOSSILS.

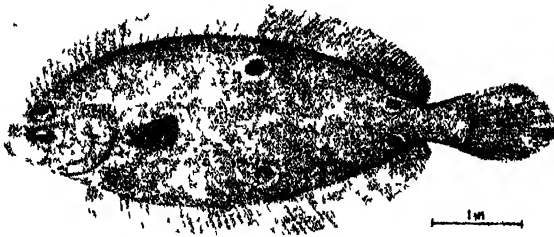
[A. S. ROMER]

### Pleuromeiales

An order of Early Triassic lycopods consisting of the genus *Pleuromeia*. The upright unbranched stem of the plant was about 1 meter tall with grass-like leaves and a single terminal strobilus of overlapping, scalelike, round sporophylls that bore one sporangium each. From its 4-lobed base sprang regularly spaced rootlets that resembled those of *Stigmaria*, the underground part of an arborescent Paleozoic plant. The internal structure of *Pleuromeia* is unknown. A similar but smaller plant, *Nathorstiana*, thrived during the Early Cretaceous. *Pleuromeia* and *Nathorstiana* may connect the Carboniferous *Lepidodendrales* with the herblike *Isoetes* of the Tertiary and Recent epochs. See LYCOPODINAE; PALEOBOTANY. [C. A. ARNOLD]

### Pleuronectiformes

One of the most distinctive orders of actinopterygian fishes, also called Heterosomata, which comprises the flatfishes: halibut, plaice, flounders, soles, tongue soles, and their allies. The striking feature of the group is the loss of bilateral symmetry, a characteristic of almost all vertebrates. Young flat-



Fourspot flounder, *Paralichthys oblongus*. (After G. B. Goode, *Great International Fisheries Exhibition*, London, U.S. Natl. Museum Bull. 27, 1883)

fishes are symmetrical and swim upright; early in life, however, one eye migrates across the top of the skull to lie on the same side as the other eye. This transformation is associated with deformation of the skull bones and nerves, a change in position so that the fish lies on one (the blind) side, partial or complete depigmentation of the blind surface, and sometimes modification and development of asymmetry in paired fins, dentition, squamation, visceral anatomy, and other structures. The dorsal and anal fins are usually long and may be confluent with the caudal; the body cavity is much restricted in size. Most species are characteristically right or left sided, but there are occasional reversed examples. Other structural characters reveal that the flatfishes are modified from perciform ancestry.

The known fossil history of the order dates from middle Eocene, when at least 2 families were well established. Recent flatfishes are classified in 6 families, about 116 genera, and nearly 500 species. They are essentially benthic inhabitants of continental shore waters, but a few ascend rivers or are found in the deep seas. Although most are tropical to temperate, a few cross the Arctic Circle. Flatfishes are of great economic importance and they rank high in the quality of their flesh. See ACTINOPTERYGII. [R.M.B.]

*Bibliography:* J. R. Norman, *A Systematic Monograph of the Flatfishes (Heterosomata)*, 1934.

### Pleuropneumonia-like organism (PPLO)

The nature of these organisms and their classification are uncertain. They may represent a special growth form of bacteria, which they resemble in many respects and, as such, have been placed in the order Mycoplasmatales. A colony is shown in Fig. 1. They include the smallest organisms capa-



Fig. 1. Colonies of a rat PPLO on agar. (From L. Dienes and G. Edsall, *Proc. Soc. Exp. Biol. Med.*, 36:740-744, 1937)

Fig. 2. Electron micrograph of a goat strain PPLO. (From E. Klieneberger-Nobel and F. W. Cuckow, *J. Gen. Microbiol.*, 12:95-99, 1955)

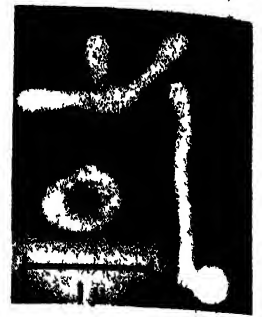
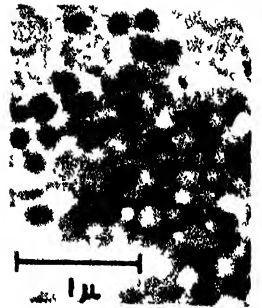


Fig. 3. Electron micrograph of a human strain PPLO. (From L. Dienes, *J. Bacteriology*, 66:280-283, 1953)

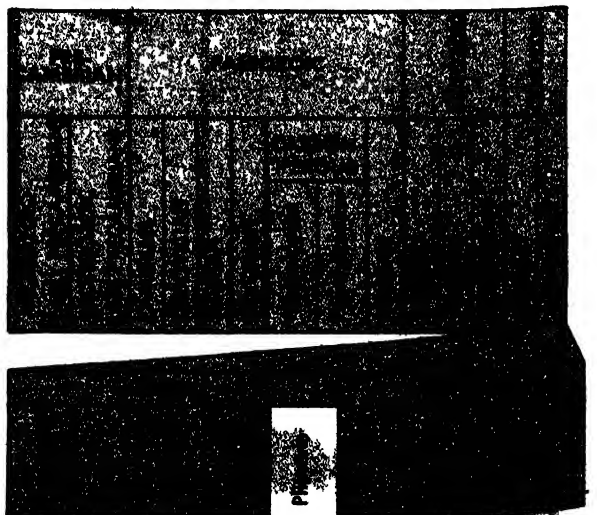


ble of independent life and are comparable in size to the larger filterable viruses (Figs. 2 and 3). The name is derived from a disease of cattle, contagious pleuropneumonia, in which the first organism of this type was discovered. Serious diseases are also produced by organisms of this group in sheep, goats, chickens, and turkeys. As harmless saprophytes, they are often present on the mucous membranes of animals and man and have been found in sewage and soil. See MYCOPLASMATALES. [L.D.]

### Pliocene

The youngest of the five major world-wide divisions (epochs) of the Tertiary Period (Cenozoic Era), the epoch of geologic time extending from the end of the Miocene to the beginning of the Pleistocene or of the Quaternary. See CENOZOIC; TERTIARY

The term Pliocene was originally proposed in 1833 by the British geologist Sir Charles Lyell who divided the Tertiary into Pliocene (youngest),



Miocene, and Eocene, and subdivided Pliocene into Newer and Older Pliocene. The latter terms were later abandoned, and Pliocene is now generally applied to what Lyell called Older Pliocene and the older part of his Newer Pliocene.

The Pliocene Series includes all rocks formed during the Pliocene Epoch, but the term is used most specifically with reference to the sedimentary rocks which were formed during this interval of geologic time. They contain the plant and animal remains which are the primary bases for identification of Pliocene age.

**Strata.** The Pliocene strata include all the common sedimentary types, both marine and continental, as well as intermediate ones. They are typically unconsolidated to poorly consolidated and are present in many parts of the world. Particularly noteworthy are the Pliocene strata of (1) the Pacific border basins of western North America; (2) the North Sea area of northwestern Europe, the Mediterranean Sea area of southern Europe and northern Africa, and the intracontinental basins of central and eastern Europe and southern Asia; (3) the Siwalik region of the Himalaya Mountains; and (4) the coastal region of South Australia. Igneous rocks are best preserved in mountainous areas, terrestrial strata are best known in the continental interiors, and marine beds are most widespread on the continental margins in the areas of the coastal plain and continental shelves. Most of the marine or marginal Pliocene strata are relatively flat-lying and occur near sea level, but crustal disturbances have appreciably deformed some of the beds and have elevated them to various heights, as for example in the Apennines of Italy and in the Pacific Border region of North America from Baja California to Alaska. Up to 5000 meters of Pliocene strata are present in parts of this Pacific Border region, but in general the thickness of these beds is much less. They contain in places, oil and gas, ground water, phosphates, sand, clay, limestone, and other products.

Pliocene strata were first studied in detail in the Mediterranean region and this is their type locality the area to which reference is made for comparing rocks in other parts of the world that are believed to be of the same age. Here the Pliocene is ordinarily subdivided into two units (stages), each of which exhibits a marine and a terrestrial facies.

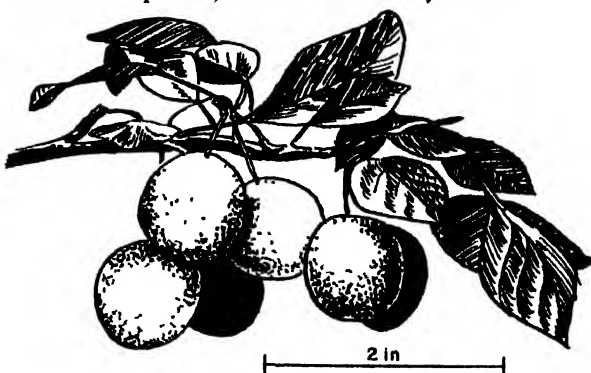
Maurice Gignoux, the modern European stratigrapher, stated in reference to the Pliocene of the Mediterranean area that no other stratigraphic unit of the Tertiary is as clean-cut; that, following the great regression at the end of the Miocene, the Pliocene formations of the region always lie in discontinuity over the older formations; and that the end of the period is always marked by a retreat of the sea. Thus, enclosed between a transgression and a regression, the Pliocene of the Mediterranean area corresponds to a sedimentary cycle, or depositional series, followed by a period of erosion which inaugurated the Quaternary. Though the Pliocene is limited in places by physical breaks in the sedimentary record, no physical

change is known at other places, and differentiation of the units is based strictly on paleontological evidence.

**Fauna.** The Pliocene fauna is distinctly modern; nearly every living type of mammal, echinoid, pelecypod, gastropod, and foraminifer is represented. In addition, many forms now extinct (saber-toothed cats, ground sloths, and glyptodonts) made up a prominent element in the Pliocene fauna. See PALEOBOTANY; PALEONTOLOGY. [A.H.CH.; G.E.M.]

## Plum

Plums are shrubs or small trees which bear smooth-skinned stone fruits also called plums. The plum is an ancient species, several thousand years old.



Damson plum. (From L. H. Bailey, *The Standard Cyclopaedia of Horticulture*, vol. 3, Macmillan, 1937)

**Varieties.** The plum includes a wide range and variety in origin, tree, and fruit. There are four principal groups: (1) *Domestica* (*Prunus domestica*) of European origin, (2) *Insititia* or Damson plums (*P. insititia*) of European origin, (3) *Salicina* (*P. salicina*) of Japanese origin, and (4) American plums (*P. americana* and *P. hortulana*). The *Domesticas* are large, meaty, prune-type plums including Agen (California prune), Reine Claude, Italian Prune, and Yellow Egg, and constitute the principal fresh and dried prunes of commerce. A prune is a plum which will dry without spoiling. *Insititias*, represented by the Shropshire and French varieties, are small, meaty fruits grown sparingly for jam. Japanese plums are typically round, reddish or yellow, and juicy, and are represented by such varieties as Abundance, Burbank, Beauty, Santa Rosa, Satsuma, and Shiro. American plums are small, watery fruits of low quality, represented by DeSoto and Pottawattomie, and valued chiefly for hardness of the tree.

**Propagation and commercial production.** Plums are propagated by budding on seedlings of the myrobalan plum (*Prunus cerasifera*), and less commonly on the peach and certain strains of *Prunus domestica*. See GRAFTING OF PLANTS; PEACH.

The *Domestica* and *Insititia* plums are slightly harder than the peach, the Japanese plums have about the same hardness as the peach, and native American plums are considerably harder than the other varieties.



Trees of the *Domestica* and *Damson* types grow to 20–30 ft, are adapted to relatively heavy soil, and are planted 20–24 ft apart. Trees of the Japanese and American types grow to 15 or 20 ft, are adapted to lighter soils, and are planted 18 ft apart. Most plums require cross pollination (see REPRODUCTION, PLANT).

Commercial production in the United States in 1957 was 88,000 tons of Japanese-type plums, of which California produced 85,000 tons, and 493,000 tons of prune-type plums of which California produced one-half. Returns to growers were \$197 a ton for plums and \$88 a ton for prunes. See FRUIT (BOTANY); FRUIT (TREE). [H.B.T.]

**Plum and prune diseases.** Brown rot, a fungus disease caused by *Monilinia fructicola*, is the limiting factor in production of plums under humid conditions. Infected fruit decays rapidly and is covered with gray masses of fungus spores (see FUNGI). Punctures made by the plum curculio, a common fruit insect, increase the risk of fungus infection. Control is difficult and must include measures for control of the plum curculio as well as the fungus. Commercial production of plums and prunes is confined almost entirely to the Pacific Coast states where the plum curculio does not occur and climatic conditions are less favorable for development of the fungus.

Japanese varieties of plums and their hybrids are so susceptible to attacks of the bacterial spot organism (*Xanthomonas pruni*) that their commercial production is likewise confined to the far western states where the organism does not occur. See BACTERIA.

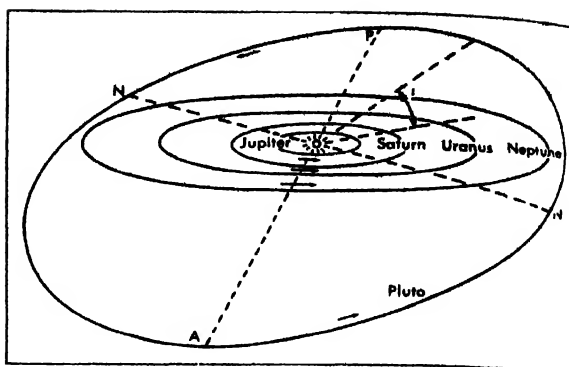
Other fungus diseases are plum pockets (a form of the leaf curl disease), rust, and scab. These diseases usually cause little damage.

Diamond canker and prune dwarf are two important virus diseases that cause prune trees to become unproductive. Control involves removal of affected trees and use of virus-free nursery stock. See FRUIT (TREE) DISEASES; PLANT DISEASE CONTROL; PLANT VIRUS. [J.C.DU.]

## Pluto

The most distant known planet in the solar system. It was discovered photographically on January 23, 1930, by C. W. Tombaugh at the Lowell Observatory, where a systematic search for a trans-Neptunian planet had been initiated by P. Lowell. The presence of an unknown planet beyond Neptune and perturbing its motion had been predicted independently by P. Lowell and by W. H. Pickering on the basis of an analysis similar to, though less rigorous than, that which had led U. J. Leverrier and J. C. Adams to the prediction of Neptune. Pluto was found surprisingly near the predicted position. However, because of the estimated low mass of the planet, some astronomers have questioned the validity of the original prediction.

Pluto's orbit has a semimajor axis (mean distance to Sun) of  $3.7 \times 10^9$  miles; an eccentricity of 0.25, the largest of the major planets; an inclination of orbital plane to ecliptic of  $17.3^\circ$ , also the



Orbit of Pluto. A perspective view to show the inclination  $i$  and eccentricity of the orbit.  $A$ , aphelion,  $P$ , perihelion,  $NN'$ , line of nodes. (From L. Rudaux and G. de Vaucouleurs, *Larousse Encyclopedia of Astronomy*, Prometheus Press, 1959)

largest of the major planets; a sidereal revolution period of 248.4 years; and a mean orbital velocity of 2.96 mi/sec. Pluto's large eccentricity causes its distance to the Sun to vary from  $4.59 \times 10^9$  miles at aphelion to  $2.76 \times 10^9$  miles at perihelion. The perihelion is  $4.9 \times 10^7$  miles less than Neptune's aphelion distance, but because of the large inclination of Pluto's orbit, the two orbits do not intersect. See NEPTUNE; PLANET.

Pluto is visible only through large telescopes, its visual magnitude at mean opposition (that is, when closest to Earth) is 14.7. The apparent diameter of its disk as estimated visually by G. P. Kuiper with the 200-in. telescope at Mount Palomar is about  $0.2''$ , and the corresponding linear diameter about half of Earth's diameter. With such a diameter and the mass, about 0.9 (Earth = 1), derived from the perturbations of Neptune, the mean density would be of the order of 40, an improbably high value. On the other hand if the albedo is about 0.2, a plausible value, the diameter would be about the same as Earth's and the density about 5, but the apparent diameter should be  $0.45''$ . This dilemma had not been resolved through 1959.

The theoretical average temperature is about  $90^\circ\text{K}$ . The spectrum of Pluto shows no trace of specific absorption attributable to an atmosphere, in particular no indication of methane, although the superficial gravity would probably be great enough to retain a tenuous atmosphere of heavy gases.

The light variations indicate that the surface of Pluto possesses dark markings and that the rotation period is about 16 hours.

Pluto has no known satellite. Its unique orbit, strongly inclined and highly eccentric, suggests an unusual origin, for example, that Pluto may once have been a distant satellite of Neptune which escaped through the effect of external perturbations on the system.

The discrepancy between the observed apparent diameter and the estimated mass of Pluto has led to the suspicion that a trans-Plutonian planet remains to be found in the outskirts of the solar system, but it would be very distant and very faint.

and the chances of finding it accidentally are extremely slight. No systematic search for such a planet had been made through 1959. [C.D.V.]

**Bibliography:** H. N. Russell, R. S. Dugan, and J. Q. Stewart, *Astronomy*, vol. 1, rev. ed., 1945.

## Pluton

A general term in geology for a rock body formed by consolidation from a magma (molten material) without reaching the surface of the earth; or possibly by processes of replacement (granitization) giving an end product not readily distinguishable from that of the first-named mechanism. As now used, the term includes all intrusive bodies regardless of size or form but does not include extrusive igneous bodies (formed at the surface of the earth). The term, first proposed by H. Cloos in 1928, had been widely adopted by geologists. See GRANITIZATION.

Plutons are concordant or discordant, depending on whether they are parallel or nonparallel to the dominant features of the rocks they intrude. The distinction has to be made with due respect to scale; concordance in a broad way is generally accompanied by discordance in detail. Moreover, a pluton obviously can be concordant over parts of its intrusive path and discordant over other parts.

**Concordant plutons.** Intrusive bodies of this type generally are divided into sills, laccoliths, lopoliths, and phacoliths, but the classification is not necessarily complete. Other forms have been described, but the names are not widely used; other form names may be added in the future, or some (such as phacolith) may be discarded.

**Sill.** A sill is a pluton relatively narrow in width but extensive in the other two dimensions, intruded essentially parallel to the planar features of the enclosed rocks. Sills generally were emplaced along the bedding planes of sedimentary rocks (Fig. 1a).

**Laccolith.** This form is commonly taken to be a cistern-shaped body with convex roof and flat floor, intruded into bedded rocks (Fig. 1b). The layers of rock over the blisterlike intrusion are pushed upward to form a dome. When eroded, the upturned edges of the resistant rock layers may form circular ridges called hogbacks. See LACCOLITH.

**Lopolith.** A lopolith is a large saucer-shaped body characterized particularly by concave roof and floor, the latter feature forming concurrent with the process of emplacement, not subsequently and independently (Fig. 1c). Though less abundant than the preceding groups, a number of large and striking examples are known.

**Phacolith.** Phacoliths are structurally the reverse of lopoliths; both roof and floor are convex upward (Fig. 1d). However, phacoliths are generally small compared to lopoliths. By definition, the shape was established concurrent with and presumably related to the process of intrusion; a sill subsequently deformed into this shape would not be a phacolith. Examples of phacoliths may be too few to warrant inclusion in a general classification.

**Discordant plutons.** Currently used names include dikes, stocks, batholiths, volcanic vents, and ring-dikes. As with concordant plutons, the list is not complete; other names have been used, and some could be added or subtracted.

**Dike.** A dike is similar to a sill in origin and, in a sense, appearance; the thickness is small compared to the extent in the other dimensions, but the body transgresses (cuts across) the features of the enclosing rocks (Fig. 2a). Dikes generally were emplaced along fractures or systems of fractures.

**Stock.** A stock is a pluton with roughly equidimensional cross section of limited size but a much greater, generally indeterminate, third dimension which ordinarily is steep or vertical (Fig. 2b).

**Batholith.** This form is most simply described as a large stock, the separation being placed at about

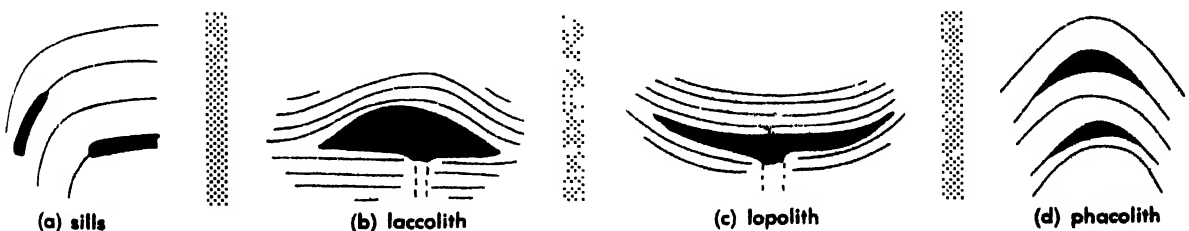


Fig. 1. (a-d) Cross sections of concordant plutons.

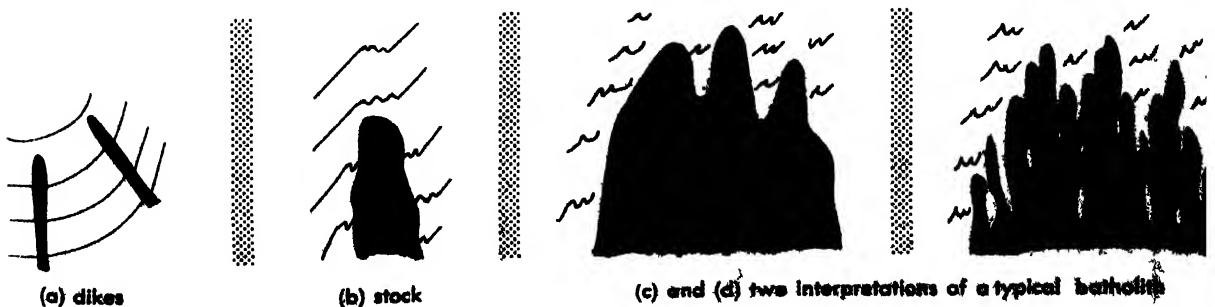


Fig. 2. (a-d) Cross sections of discordant plutons.

40 mi<sup>2</sup> of cross-sectional area, but a widely used definition (by R. A. Daly, 1933) also stated for both stocks and batholiths that no floor could be determined or inferred, and this usage is preferred by some geologists (Fig. 2c and d). See BATHOLITH.

**Volcanic vents.** These intrusions are the roots of volcanoes exposed by erosion or by underground explorations (see VOLCANO). They are generally composites of several intrusive and extrusive features and are complex.

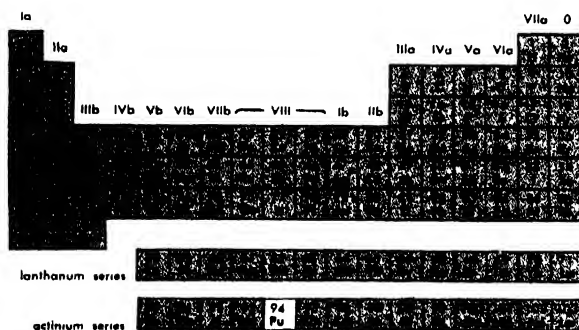
**Ring-dikes.** Ring-dikes are arcuate dikes formed under special circumstances during the emplacement of stocks or batholiths; they are known from a few widely separated places in the world.

[J.A.N.]

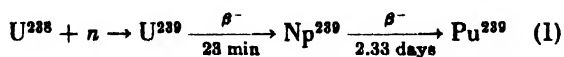
**Bibliography:** M. P. Billings, *Structural Geology*, 2d ed., 1954; R. A. Daly, *Igneous Rocks and the Depths of the Earth*, 1933; C. B. Hunt, P. Averitt, and R. L. Miller, *Geology and Geography of the Henry Mountains Region, Utah*, USGS Profess. Paper 228, 1953.

## Plutonium

A chemical element, Pu, atomic number 94. Plutonium is a reactive, silvery metal in the transuranium series of elements. The first isotope to be identified was Pu<sup>238</sup>, produced in cyclotron experiments by



G. T. Seaborg, E. M. McMillan, A. C. Wahl, and J. Kennedy. The principal isotope of chemical interest is Pu<sup>239</sup>. It is formed in nuclear reactors by the process



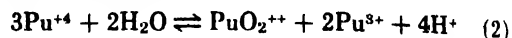
Pu<sup>239</sup> decays by  $\alpha$ -emission with a half-life of 24,360 years. Its fissionability makes it of importance in nuclear weapons and in nuclear reactors. Minute quantities of Pu<sup>239</sup> are formed in pitchblende, and monazite ores by reaction (1). In pitchblende, the uranium to plutonium ratio is approximately 10<sup>11</sup>:1.

**Uses.** Plutonium is used as a nuclear fuel, to produce radioactive isotopes for research, and as the fissile agent in nuclear weapons. See NUCLEAR FUELS; REACTOR, NUCLEAR; REACTOR, NUCLEAR (CLASSIFICATION).

**Properties.** Like its neighboring elements, uranium and neptunium, plutonium exhibits a variety of valence states in solution and in the solid state. In solution, the known oxidation states are III, IV,

V, and VI. Plutonium metal is highly electropositive. The ions of the IV, V, and VI states are moderately strong oxidizing agents. The ions of the II, IV, and VI states can coexist in 1 M perchloric acid solution.

Because the oxidation potentials are so close in value, pure solutions of intermediate oxidation states undergo disproportionation (self-oxidation and reduction reactions). The most important equilibrium is that involving the disproportionation of Pu(IV), which can be written



for which the equilibrium constant is

$$K_1 = \frac{[\text{PuO}_2^{2+}][\text{Pu}^{3+}]^2[\text{H}^+]^4}{[\text{Pu}^{4+}]^3} \quad (3)$$

K<sub>1</sub> is calculated from the potentials to be 0.0089 for 1 M acid at 25°C. In 1 M acid, the solution resulting from the disproportionation of pure plutonium(IV) would be 72% Pu(IV), 18.6% Pu(III), and 9.3% Pu(VI). Although Pu(V) is unstable in molar perchloric acid, it becomes increasingly stable as the acidity decreases. In 0.1 M acid, appreciable concentration of all four valence states may coexist in solution. The additional equilibrium that must be considered may be written as

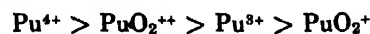


The equilibrium constant, K<sub>2</sub>

$$K_2 = \frac{[\text{PuO}_2^{2+}][\text{Pu}^{3+}]}{[\text{PuO}_2^+][\text{Pu}^{4+}]} \quad (5)$$

is 13 in 1 M perchlorate solution at 25°C. The rate of attainment of the equilibrium of reaction (2) is slow and that of reaction (5) very fast. A constant equilibrium state is not maintained over long periods of time because of the slow reduction in average oxidation number in solution caused by the reaction of the plutonium ions with the  $\alpha$ -radiation-induced decomposition products of the water.

In solutions of acids, such as nitric and hydrochloric, whose anions form weak complexes with plutonium ions, the relative stabilities of the different states are little changed. Qualitatively, it is known that univalent anions, with the exception of fluoride, form relatively weak complexes with the ions of all oxidation states. Higher-valent anions form relatively strong complexes. In general, the relative stabilities of complexes with a given anion decreases in the order



Complex formation will generally stabilize the IV state. Hydrolysis reactions can also markedly affect the relative stabilities of the different states. As in the case of complex formation, the IV state is stabilized by hydrolysis. Polymerization processes are important in the hydrolysis reactions. Plutonium(IV) has been reported to form soluble polymers with molecular weights as high as 10<sup>10</sup>.

A large amount of information is available on the behavior of plutonium ions when treated with

common oxidizing and reducing agents. It is generally found that reactions which involve only changes from the III to IV or V to VI states tend to be rapid. Reactions which involve the formation or destruction of the oxygenated ions of the V or VI states tend to be slow. As examples, oxidation of Pu(III) to Pu(IV) is rapid with bismuthate, bromate, iron(III), dichromate, iodate, permanganate, and cerium(IV) ions. Reduction of Pu(IV) to Pu(III) is rapid with iron(II), iodide ion, sulfurous acid, and nitrous acid. Oxidation of Pu(IV) to Pu(VI) is slow with bromate, dichromate, permanganate, and nitrate ions. The rates may be changed markedly by complex-ion formation. In the presence of moderate concentrations of sulfuric acid, for example, oxidation past the IV state is very difficult. Some relatively rapid oxidation reactions are also known. Ceric ion, argentic ion, and bismuthate rapidly oxidize Pu(IV) to Pu(VI).

The ions of the different oxidation states have characteristic colors: Pu<sup>3+</sup> is blue-violet; Pu<sup>4+</sup>, yellow-brown; PuO<sub>2</sub><sup>+</sup>, reddish; and PuO<sub>2</sub><sup>2+</sup>, pink like the rare earths, they also have characteristic absorption spectra with sharp absorption bands. These have been widely used in the analysis of plutonium solutions to determine the amount of each oxidation state present.

**Preparation of the element.** Methods for the isolation and purification of plutonium make use of the fact that the element can exist in a multiplicity of oxidation states, each differing in chemical properties. Laboratory separation procedures have been devised using carrier, solvent-extraction, and ion-exchange methods. The first plant-scale operations employed the carriers bismuth phosphate and lanthanum fluoride. In most recent processes, solvent extraction is employed. It has the advantage that not only the plutonium but also the uranium of the reactor fuel may be readily recovered and decontaminated from fission products. Some of the most important solvents are listed in Table 1. Control of the extraction behavior is obtained by the use of diluents for the solvent, addition of salting agents to the aqueous layer, and the control of solution pH. In Table 1 are listed the diluents and salting agents commonly employed for the different solvents. The behavior of different valence states with these solvents can be illustrated by reference to the relative distribution coefficients, defined as the concentration of the metal in the organic phase divided by the concentration in the aqueous phase (Table 2). The actual values of the distribution coefficients will change with conditions, but the approximate relative values will be maintained.

In the industrial process employing hexone (the Redox process), the uranium fuel is dissolved in nitric acid. The solution is oxidized, and the U(VI) and Pu(VI) are coextracted from the fission products. After scrubbing the hexone layer to remove impurities, the solvent is passed over an aluminum nitrate solution containing a reducing agent. The plutonium is removed into the aqueous layer as Pu(III), and the uranium left in the solvent as U(VI). The aqueous layer is then reoxidized and

Table 1. Solvents used in the separation of plutonium and uranium from fission products

Solvent (trivial name)	Diluent	Salting agent
Methyl isobutyl ketone (Hexone)	None	Al(NO <sub>3</sub> ) <sub>3</sub>
Tri- <i>n</i> -butyl phosphate (TBP)	Kerosine	HNO <sub>3</sub> , Al(NO <sub>3</sub> ) <sub>3</sub>
Dibutyl ether of ethylene glycol (Carbitol)	None	HNO <sub>3</sub>
Dibutyl ether of tetra-ethylene glycol (Pent-ether)	None or butyl ether	HNO <sub>3</sub>
Triglycol dichloride (Trigly)	None	Al(NO <sub>3</sub> ) <sub>3</sub>
Thenoyl trifluoroacetone (TTA)	Benzene or toluene	None

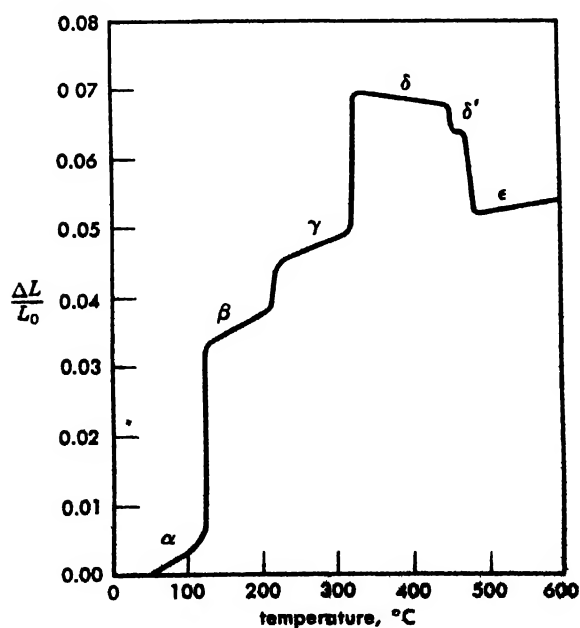
Table 2. Distribution coefficients of uranium, plutonium, and fission products from nitrate solutions of 100-day cooled reactor fuel

Solvent	U(VI)	Pu(VI)	Pu(IV)	Pu(III)	Fission products
Hexone	15	76	$1.6 \times 10^{-3}$	$4.5 \times 10^{-4}$	$6 \times 10^{-4}$
TBP	80	0.6	1.5	$2 \times 10^{-3}$	$2 \times 10^{-3}$

the extraction repeated. By successive cycles, the plutonium is purified to the desired degree.

The industrial process employing tri-*n*-butyl phosphate (TBP) as the solvent (the Purex process) operates in much the same manner. After dissolution of the fuel element, the plutonium is fixed as Pu(IV) and the uranium as U(VI). The nitric acid concentration is adjusted, and the Pu(V) and U(VI) extracted into 30% TBP in kerosine. The solvent is washed with nitric acid to remove impurities. The plutonium is then removed as Pu(III) by scrubbing the solvent with nitric acid containing a reducing agent.

Plutonium metal can be prepared by the reduction of PuF<sub>3</sub> with calcium metal. Plutonium metal



Expansion of high-purity plutonium under conditions of self-heating,  $L_0 = 0.5$  in. (After E. R. Jette)

Table 3. Properties of plutonium metal

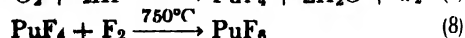
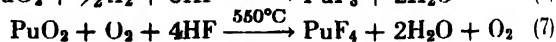
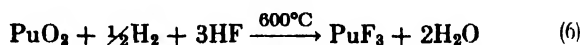
Phase	Symmetry	Density (20°C), g/cm <sup>3</sup>	Temperature of phase transition, °C	Linear expansion coefficient* (20°C) $\alpha \times 10^6$	Resistivity (20°C) $\times 10^6$ (ohm-cm)	Temperature coefficient of resistivity†
$\alpha$	Monoclinic	19.82		50.8	145	-21
$\beta$		17.65	122	38.0	110.5	-6
$\gamma$		17.19	203	34.7	110	-5
		17.14				
$\delta$	Face-centered cubic	15.92	319	-10.0	103	+7
$\delta'$	Body-centered tetragonal	16.0	453	-20	105	+45
			477			
$\epsilon$	Body-centered cubic	16.48	639.5	25.7	114	-7
Liquid		16.5		50		

$$* \alpha = \frac{1}{L} \cdot \frac{L}{T} \quad \dagger \frac{1}{\rho} \cdot \frac{\rho}{t} \times 10^6$$

deserves special mention because of its unique properties. It is known to exist in six allotropic forms below the melting point (639°C). Some of its physical properties are given in Table 3. Particularly interesting and puzzling are the contractions which the  $\delta$  and  $\delta'$  phases undergo with increasing temperature (see graph); noteworthy is the fact that for no phase do both the coefficient of thermal expansion and the temperature coefficient of resistivity have the conventional algebraic sign. If the phase expands on heating, the resistance decreases. A number of alloys of plutonium are known—with beryllium, lead, uranium, chromium, manganese, iron, nickel, and osmium.

**Principal compounds.** A large number of compounds of plutonium have been prepared. Reaction

of hydrogen with plutonium metal yields at least two well-defined hydrides, PuH<sub>3</sub> and PuH<sub>2</sub>. The common oxide is PuO<sub>2</sub>. It is formed by ignition of the hydroxides, oxalates, peroxides, and nitrates of any oxidation state in air at 870–1200°C. It crystallizes in a face-centered-cubic structure (density 11.44 g/cm<sup>3</sup>). It has been extensively used for gravimetric analyses of plutonium. A lower oxide Pu<sub>2</sub>O<sub>3</sub> is known. One of the most important classes of compounds is made up of the halides. Properties of the known halides and oxyhalides are given in Table 4. The hexafluoride is a low-melting, low boiling compound of high volatility resembling NpF<sub>6</sub> and PuF<sub>6</sub>. It is a strong fluorinating agent. Conditions for the preparation of the fluorides are illustrated by the equations



The other halides are prepared by a variety of methods. Treatment of PuO<sub>2</sub> with powerful halogenating agents such as CCl<sub>4</sub>, PCl<sub>5</sub>, and SCl<sub>2</sub> yields PuCl<sub>3</sub>, PuBr<sub>3</sub> and PuI<sub>3</sub> are conveniently made by the action of the anhydrous gases, HBr and HI, on plutonium metal.

A number of other compounds are known. Among these are the carbides, silicides, sulfides, and nitrides. These are of interest because of their

Table 4. Plutonium halides and oxyhalides

Compound	Color	Melting point, °C	Density at 20°C	Crystal structure
PuF <sub>3</sub>	Purple	1425	9.32	Hexagonal
PuF <sub>4</sub>	Pale brown	1037	7.0	Monoclinic
PuF <sub>5</sub>	Reddish-brown	50–75		Orthorhombic
PuCl <sub>3</sub>	Green	760	5.70	Hexagonal
PuBr <sub>3</sub>	Green	681	6.69	Orthorhombic
PuI <sub>3</sub>	Green	777	6.92	Orthorhombic
PuOF	Metallic	1635	9.76	Tetragonal
PuOCl	Blue-green		8.81	Tetragonal
PuOBr	Green		9.07	Tetragonal
PuOI	Green		8.46	Tetragonal

Table 5. Some insoluble inorganic compounds of plutonium precipitated from aqueous solution

Oxidation state			
III	IV	V	VI
Pu(IO <sub>3</sub> ) <sub>3</sub>	NH <sub>4</sub> PuF <sub>6</sub>	KPuO <sub>2</sub> CO <sub>3</sub>	NaPuO <sub>2</sub> (C <sub>2</sub> H <sub>3</sub> O <sub>2</sub> )
PuPO <sub>4</sub> ·0.5H <sub>2</sub> O	Pu(OH) <sub>4</sub> ·xH <sub>2</sub> O		
Pu <sub>3</sub> (C <sub>2</sub> O <sub>4</sub> ) <sub>2</sub> ·9H <sub>2</sub> O	Pu(IO <sub>3</sub> ) <sub>4</sub>		
	PuO <sub>4</sub> ·2H <sub>2</sub> O		
	Pu(HPO <sub>4</sub> ) <sub>2</sub> ·xH <sub>2</sub> O		
	Pu(C <sub>2</sub> O <sub>4</sub> ) <sub>2</sub> ·6H <sub>2</sub> O		

refractory nature. Among these are  $\text{PuC}$ ,  $\text{Pu}_2\text{C}_3$ ,  $\text{PuN}$ ,  $\alpha\text{-PuSi}_2$ ,  $\beta\text{-PuSi}_2$ ,  $\text{PuSi}$ ,  $\text{PuS}$ ,  $\text{Pu}_2\text{S}_3$ - $\text{Pu}_3\text{S}_4$ .

In addition to those compounds prepared by vacuum line techniques, there are a large number of compounds that have been prepared from solution. The most important of these are given in Table 5. See NEPTUNIUM; NUCLEAR CHEMISTRY; TRANSURANIUM ELEMENTS. [J.C.H.]

**Bibliography:** J. J. Katz and G. T. Seaborg, *The Chemistry of the Actinide Elements*, 1957; G. T. Seaborg and J. J. Katz (eds.), *The Actinide Elements*, 1954; G. T. Seaborg, J. J. Katz, and W. M. Manning (eds.), *The Transuranium Elements*, 1950.

## Plywood

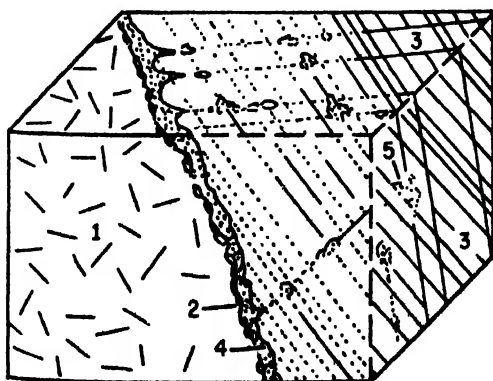
Glued sheets, strips, or pieces made of three or more cross-grained wood veneers (thin plywoods), or two, four, or more veneers plus a sawed lumber or multiple-block core (core plywoods). The cross-graining of the plies gives plywood much greater strength and dimensional stability than the plain wood from which the plies are made.

Thin hardwood veneers over less costly softwood cores are widely used in furniture and for wall and ceiling panels. Plywoods made of inexpensive woods are employed extensively in concrete forms, sheathing, subflooring, boxes, and crates. Molded, die-pressed, or bag-molded plywoods are increasingly important in the manufacture of small boat hulls, aircraft parts, chairs, and similar shell-like constructions. See LUMBER MANUFACTURE. [G.CO.]

## Pneumatolysis

The alteration of rocks by the action of magmatic gases. The gases accompany magmatic intrusions and permeate the intruded rocks along fissures and other lines of least resistance. See METAMORPHISM; METASOMATISM.

Primary magmatic gases are acid and therefore react readily with limestone to form skarn rocks.



Skarn rocks and pneumatolytic ore at Aranzazu, Mexico: 1, granodiorite; 2, garnetized border of granodiorite (exaggerated); 3, limestone; 4, garnet rock at immediate contact carrying some ore; 5, bodies of andradite-wollastonite-copper ore localized along intersections of fissures and bedding planes. (After A. Knapf, in W. H. Newhouse, *Ore Deposits as Related to Structural Features*, Princeton Univ. Press, 1942)

Appreciable amounts of heavy metals (usually present as chlorides, fluorides, or sulfides) are associated with the magmatic gases. These are captured by the limestone and retained as deposits of skarn rocks or ores. In other places the heavy metals are deposited in granite or schist, causing the formation of greisen. See GREISEN; SKARN.

The following list indicates common pneumatolytic minerals concentrated in limestone of the Oslo region.

Metals	Minerals
Fe	Andradite, hedenbergite, oxidic and sulfidic iron ore
Zn, Cu, Pb	Sphalerite, $\text{ZnS}$ ; chalcopyrite, $\text{CuFeS}_2$ ; galena, $\text{PbS}$
Mn	Andradite; hedenbergite; rhodonite, $(\text{Mn,Fe,Ca})\text{SiO}_3$
Bi, Ag	Bismuthinite, $\text{Bi}_2\text{S}_3$ ; galena; sphalerite
Mo, W	Molybdenite, $\text{MoS}_2$ ; scheelite, $\text{CaWO}_4$
Co, As, Sb	Cobaltite, $\text{CoAsS}$ ; arsenopyrite, $\text{FeAsS}$ ; bismuthinite
Be, Ce	Helvite, $(\text{Mn,Fe,Zn})_4\text{Be}_3\text{Si}_2\text{O}_{12}\text{S}$ ; vesuvianite; allanite, $(\text{Ca,Fe})_2(\text{CeAl,Fe})_3\text{Si}_3\text{O}_{12}\text{OH}$
Metalloids	
Si	Silicates in skarn, quartz
F, Cl, S	Fluorite, $\text{CaF}_2$ ; scapolite; sulfidic ore
B, P, Ti	Axinite, apatite, sphene

Additional pneumatolytic minerals are those of the humite group and, by pneumatolysis of shale, tourmaline, topaz, muscovite, phlogopite, and lithium micas. Luxullianite is a tourmalinized granite.

In conclusion it may be emphasized that limestone is especially susceptible to pneumatolytic contact metamorphism, resulting in formation of skarn rocks and often useful ore deposits (Iron Springs, Utah; Macay, Idaho; Yerrington, Nevada; Concepción del Oro, Mexico). The minerals often develop as well-formed crystals, and the deposits belong to the best known mineral occurrences in the world (Franklin Furnace, N.J.; Clifton-Morenci, Ariz.; Auerbach, Germany; Berggieshübel; Banat, Hungary; Concepción del Oro, Mexico). Silicate rocks are not usually as intensively altered, but by introduction of fluorine and lithium, and by the formation of greisen, feldspars may change into topaz, zinnwaldite, or other micas. Tin is sometimes introduced, forming cassiterite,  $\text{SnO}_2$ . Other introduced minerals in argillites are, for instance, molybdenite, apatite, or beryl.

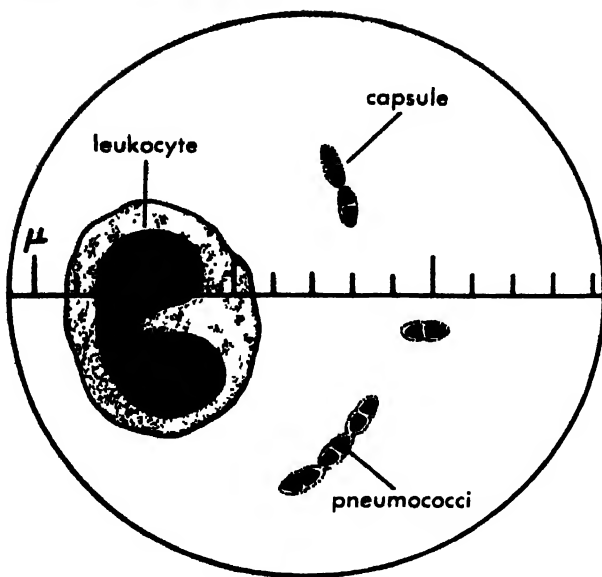
Sometimes intense boron pneumatolysis may produce datolite, axinite, and rare minerals like kotoite,  $\text{Mg}_3(\text{BO}_3)_2$ ; fluoborite,  $\text{Mg}_3(\text{BO}_3)(\text{F,OH})_2$ ; ludwigite,  $(\text{Mg,Fe})_2\text{Fe}^{III}\text{BO}_5$ ; and others. [T.F.W.B.]

## Pneumococcus

The major bacterial cause of lobar pneumonia. The microorganism is referred to as *Diplococcus pneumoniae* (tribe Streptococcaceae of family Lactobacillaceae). See LACTOBACILLACEAE.

**Morphology.** Pneumococci occur as pairs of oval or lancet-shaped cocci, 0.5–1.25  $\mu$  each, flattened





Pneumococcus in sputum. (From A. B. Sabin, *J. Am. Med. Assoc.*, 100(20):1585, 1933)

at proximal sides and pointed at distal ends. A capsule envelops each pair or chain of cocci; this capsule may have a uniform periphery or may show indentations between the twin cells or between pairs. The organism is nonmotile and stains gram-positive unless the organisms are degenerating or dead. See BACTERIAL MOTILITY; GRAM'S STAIN.

**Cultural characteristics.** Pneumococci are fastidious in their nutritional requirements, which include many amino acids, vitamins, minerals, and carbon sources. The organism grows best in enriched media containing serum or blood and in the presence of oxygen, being aerobic. It also may grow anaerobically, in which case it probably derives oxygen through a flavin-containing enzyme system. It produces hydrogen peroxide that may hinder or prevent growth; this may be overcome by adding substances having oxidation-reduction action. It grows uniformly in liquid media. Lactic acid and, in some strains, formic and acetic acids, accumulate in the culture medium, increasing the acidity and limiting growth; this may be overcome by adding glucose to the medium and by neutralizing the acid with sodium hydroxide. Growth is optimum at 37°C and the organisms die rapidly at 55°C.

On blood agar, the organism produces small, water-clear, flattened, and (later) umbiliform colonies that are surrounded by a greenish zone of hemolysis, indicating methemoglobin formation. The organism contains intracellular proteolytic, lipolytic, and carbohydrate-fermenting enzymes; these readily produce autolysis, that is, dissolution of the organism. Viability of the pneumococcus can be preserved in partly desiccated animal tissue, like mouse heart or spleen, or by lyophilization, which is rapid freezing with vacuum drying. See LYOPHILIZATION.

**Identification.** Pneumococcus closely resembles many species of *Streptococcus viridans*, but can be reliably differentiated by either of the following methods.

**Solubility in bile.** The addition of a solution of either bile salts or sodium deoxycholate to a young broth culture or to suspensions of organisms, at a neutral pH, will cause lysis of the pneumococcus, and a turbid pneumococcal suspension will clear rapidly. This action is complete in 5–10 min and probably represents accelerated autolysis. It occurs without regard to specific type or colonial form. *S. viridians* will not lyse under this test.

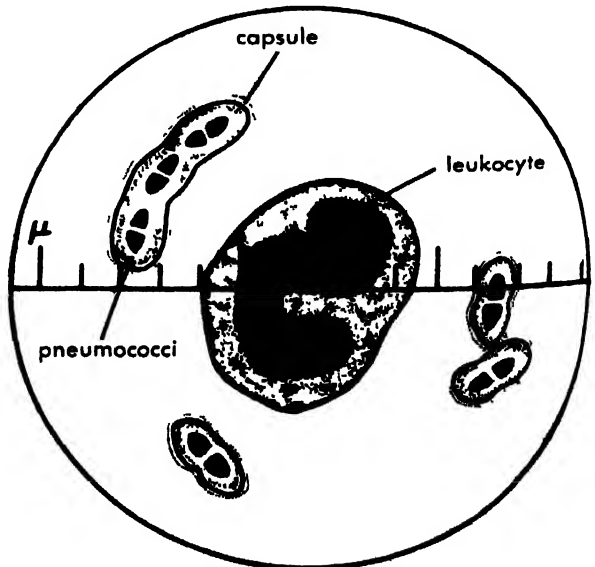
**Quellung reaction.** Swelling of the pneumococcus capsule is induced by mixing the type-specific antiserum (serum containing antibodies) prepared from rabbits, with a suspension of the organisms. This differentiates the type as well as the species. See QUELLUNG REACTION.

**Pneumococcal agents.** Pneumococci are vulnerable to the usual germicides and antiseptics, growth is inhibited by many quinine derivatives, especially optochin (ethyl hydrocupreine), and also by most sulfonamide drugs (see ANTISEPTIC; SULFA DRUGS). They are susceptible in vitro and in vivo to many antibiotics; the most effective is penicillin. Although resistance to these antibacterial agents has been induced, by repeated subcultures in subinhibitory concentrations, resistance has not been shown to develop during treatment of pneumococcal infections.

**Antigens.** Pneumococci contain three major antigens. See ANTIGEN.

**Somatic protein antigen.** This antigen is common to all types of pneumococci and to certain streptococci. The individual is afforded little, if any protection by antibodies to this fraction, upon exposure to pneumococci or during the course of such an infection.

**Somatic C carbohydrate.** The somatic C carbohydrate is common to all types of pneumococci, but not to all streptococci. Skin reactions are produced when this antigen is injected intradermally into man.



Quellung reaction. Type II pneumococcus mixed with Type II rabbit antiserum. (From A. B. Sabin, *J. Am. Med. Assoc.*, 100(20):1585, 1933)

dividuals with active pneumococcal infections of all types and also with certain streptococcal or other infections. Some minor, non-type-specific protection is afforded by antibodies to this soluble antigen derived from certain strains. This factor also forms a portion of the Forssman, or heterophile, antigen.

SSS is a contraction for type-specific soluble capsular polysaccharide. As little as 0.001 mg of this material, from some types, produces antibodies in the blood that react either directly with the specific carbohydrate of the capsule or with the whole organism, thus protecting against infection by organisms of the same type. Intradermal injection of SSS into patients recovering from pneumonia of the same type give immediate (wheal and erythema) local skin reactions. Serum from patients acutely ill with pneumococcal infections also contain a C-reacting protein, apparently not an antibody, that is also found in many other infections and febrile illnesses.

**Types.** Pneumococci can be divided into more than 77 types on the basis of the differences in their highly polymerized, immunologically distinct, polysaccharides. The specificity of the types depends on their chemical composition and structure. This has been worked out for a few types. Some of these polysaccharides are closely related, immunologically and chemically, to similar ones contained in certain yeasts, fungi, the hemicelluloses of many plants, dextrans, blood group A substance, and to those from other bacterial species, notably *Klebsella pneumoniae*, *Bacillus anthracis*, and *Hemophilus influenzae*. There are also serological cross reactions between certain types, notably between II and V and between III and VIII. The specific polysaccharides are also subject to hydrolysis by enzymes produced by various soil organisms; many of the enzymes are highly type-specific.

**Variants.** In the past, descriptions of morphological variants in pneumococcal colonies were based on their appearance on the surface of solid media. The terms most used were smooth (S), rough (R), and mucoid (M). Problems arose because of the recognition of colonies intermediate between mucoid and smooth and smooth and rough. R. Austrian (1953) proposed that the definition of morphologic variant be made at a cellular rather than at a colonial level. He defined the morphologic variants grown under suitable conditions as the following: nonfilamentous, capsulated (S); nonfilamentous, noncapsulated (R); filamentous capsulated (?) and filamentous, noncapsulated (also R). The filamentous forms differ from the nonfilamentous in that the cells of the filamentous forms fail to separate after division, giving rise to chains of pneumococci where one nonfilamentous form appears as single cells, diplococci, or short chains. All the capsulated forms have specific polysaccharides; the noncapsulated do not. Rough strains, which are avirulent and lack type specificity, may be produced by growth in homologous, type-specific antiserum, that is, serum containing antibodies of the same pneumococcal type. Rough strains, as well as heat-killed strains, have been transformed into virulent smooth forms, of types other than the original one,

by growth either in solutions of the type-specific organisms or in solutions of deoxyribonucleic acid derived from the type-specific organisms. Type transformations can also be made in mice inoculated with mixtures of both dead, encapsulated organisms of one type and living, rough nonencapsulated organisms of another. Deoxyribonucleic acids have also been used to transform strains sensitive to antibiotics, such as penicillin and streptomycin, into antibiotic-resistant strains, without employing exposure to these antibiotics. Nearly all the 20 variants, theoretically deduced by different combinations of the four morphologic variants, have also been produced in this manner. This is analogous to recombination of independently heritable characters in sexually reproducing forms. See BACTERIAL GENETICS.

**Pathogenicity.** Pneumococci occur in the upper respiratory tract of apparently healthy guinea pigs, rabbits, horses, calves, dogs, monkeys, and humans. Epizootics of pneumococcal pneumonia, and of other local or systemic pneumococcal infections, occur in monkeys, guinea pigs, and rats but are not the source of human infections (see PNEUMONIA). Epidemics of pneumococcal pneumonia also occur in humans, mostly in closed institutions. Some of the epidemics have been prevented, or their spread halted, by immunization with killed vaccines or with specific carbohydrates of the same pneumococcal types. In man, pneumococci may be found in the upper respiratory tract of nearly all individuals at some time or other. Pneumococci are the principal cause of lobar pneumonia, but may also produce meningitis, pericarditis, and peritonitis as well as infections of the pleura, middle ear, and accessory nasal sinuses. Some of these infections are accompanied by invasion of the blood stream, or septicemia. The mortality in untreated infections is high, but this has been markedly reduced by the use of antibiotics. See BACTERIOLOGY, MEDICAL. [M.F.]

**Bibliography:** R. Austrian, Morphologic variation in pneumococcus, *J. Exp. Med.*, 98:21-34, 35-40, 1953; R. J. Dubos (ed.), *Bacterial and Mycotic Infections of Man*, 3d ed., 1958; M. Finland, Recent advances in the epidemiology of pneumococcal infections, *Medicine*, 21:307-344, 1942; M. Heidelberger, The formation of antibodies in man after injection of pneumococcal polysaccharides, *Proc. Natl. Acad. Sci. U.S.*, 43:883-887, 1957.

## Pneumocystosis

A disease of man caused by a sporozoon, *Pneumocystis carinii*. The microorganism, originally reported in rats in 1912, has been increasingly recognized as a widespread human pathogen since 1953, chiefly through the investigations of O. Jirovec and collaborators. Only one species is recognized in animals and man, and accordingly there is a supposition that the infection may pass from one of these groups to another. Of possible importance is the role of adult carriers in disseminating the infection to infants.

Most cases reported thus far have been fatal, occurring in infants or very young children. The lung

lesions are conspicuous and distinctive, and the entity plasma-cell interstitial pneumonia has been attributed to *Pneumocystis*. Satisfactory therapy and control remain to be achieved. See PARASITOLOGY, MEDICAL. [D.W.]

**Bibliography:** J. Vaněk, O. Jírovec, and J. Lukeš, Interstitial plasma cell pneumonia in infants, *Ann. Paediat.*, 180:1-21, 1953.

## Pneumonia

An acute or chronic inflammation of the lung tissues, occurring in humans and in many animals. Pneumonia in humans may be caused by numerous microbial, immunological, physical, or chemical agents. It may also be associated with many systemic diseases. Any or all parts of the lung may be involved, with the inflammatory exudate filling the alveolar air spaces of one or more lobes as in lobar pneumonia, or the smaller segments in lobular pneumonia. It may be disposed in and around the bronchi in bronchopneumonia or may be limited predominantly to the interalveolar areas, as in interstitial pneumonia. Pneumonia may be a primary disease or a secondary event in other diseases.

**Known causes and diseases.** Pneumonia may be caused by bacteria, fungi, various parasites, and rickettsiae and also by miscellaneous conditions like allergy, exposure to chemicals, and foreign bodies.

**Bacteria.** The following organisms cause pneumonia:

Pneumococci of various types cause pneumococcal pneumonia.

Staphylococci that are coagulase positive cause staphylococcal pneumonia.

Hemolytic group A streptococci cause streptococcal pneumonia.

*Klebsiella pneumoniae*, types A, B, and C, cause Friedlander's pneumonia.

*Bordetella pertussis* causes whooping cough, often with pneumonia as a secondary infection or complication.

*Haemophilus influenzae* causes influenzal pneumonia. Type b strains of *H. influenzae* are usually found in infants; the strains found in adults cannot be typed.

*Pasteurella pestis* causes bubonic, or pneumonic, plague.

*Brucella* species cause brucellosis. In acute brucellosis, pneumonia may occur as a complication.

Coli-aerogenes organisms may cause pneumonia, incidental to a systemic infection.

*Salmonella* species cause systemic infections, including typhoid fever. Pneumonia may occur as a complication.

Meningococci cause meningococcal meningitis, with pneumonia as a secondary infection or complication.

*Bacillus anthracis* causes anthrax. One form, pulmonary anthrax, has a high fatality rate if not diagnosed early.

*Mycobacterium tuberculosis*, on occasion, may cause acute tuberculous pneumonia.

Mixed infections, that is, infections with more than one species of bacteria, or with bacteria and viruses, may cause pneumonia. See BACTERIOLOGY, MEDICAL.

**Fungous infections.** Infections due to fungi may involve the lung and cause pneumonia in moniliasis, actinomycosis, blastomycosis, cryptococcus, histoplasmosis, coccidioidomycosis, nocardiosis, and others (see MYCOLOGY, MEDICAL).

**Parasitic diseases.** Infections with protozoa or helminthes may have pulmonary manifestations, often acute, in such diseases as trichinosis, malaria, amebiasis, Leishmaniasis (kala-azar), toxoplasmosis, schistosomiasis, paragonimiasis, clonorchiasis, and filariasis (see PARASITOLOGY, MEDICAL).

**Viruses.** The known viruses which may cause pneumonia are influenza A (including Asian type) A' and B; psittacosis (ornithosis); lymphocytic choriomeningitis; variola (smallpox); varicella (chicken pox); measles; adenovirus type 4 (and possibly other types); lymphogranuloma venereum; and feline pneumonia. Of probable but unproved viral origin are the pulmonary involvements in cytoplasmic inclusion disease of infants, infectious mononucleosis, and the so-called primary atypical pneumonia (PAP), or viral pneumonia (see ANIMAL VIRUS).

**Rickettsial infections.** These include Q fever epidemic typhus (including the recrudescence form or Brill's disease), Rocky Mountain spotted fever, *fièvre boutonneuse*, South African tick fever, and probably others. See RICKETTSIOSES.

**Miscellaneous.** The following pneumonias have less common, or ill-defined etiologic agents: Loeffler's eosinophilia (allergic pneumonia), rheumatic pneumonia, lipoid pneumonia (from lipids and oils, as oily nose drops), inhalation or aspiration pneumonia (from chemicals, gases, blood, and other foreign bodies), pulmonary hemosiderosis (associated with hemolysis of red blood cells, usually from a systemic disorder).

**Complications.** The bacterial pneumonias may sometimes cause destruction of parts of the lung by abscess formation or may involve the pleura in an empyema or sterile effusions. Inflammation and dilation of bronchi (bronchiectasis) and spread of infection to remote organs is not unusual. Fortunately, however, most common pneumonias heal by resolution, leaving lung structure and function little changed.

Viral pneumonias may have superimposed bacterial infection.

**Therapy and prognosis.** Treatment is specific for the causative agent; chief reliance is on antimicrobial agents. See ANTIBIOTIC; CHEMOTHERAPY.

The course and outcome depend on the cause and also on the availability and proper use of specific therapy. [M.F.]

**Bibliography:** R. Heffron, *Pneumonia, with Special Reference to Pneumococcus Lobar Pneumonia*, 1939; H. A. Reimann, *Pneumonia*, 1954.

## Pneumonitis

An atypical pneumonia caused by one of several large viruses of the lymphogranuloma-psittacosis group. These are basophilic agents which have been placed in a subdivision and called pneumonitis agents. They have caused pneumonitis in man, mice, cats, sheep, goats, and cattle. In some cases there is a question whether the virus resided in the inoculum being tested or in the respiratory tract of the experimental animal used for isolation. There is also a question whether those isolated from man are specifically adapted or whether they are psittacosis viruses of exceptional virulence and unknown origin. Their public health significance is not known. See LYMPHOGRANULOMA-PSITTACOSIS GROUP. [K.F.M.E.]

## Podicipitiformes

The order of birds containing the single family Podicipitidae, the grebes. Long thought to share a common ancestor with the loons (Gaviiformes), the grebes are now believed to have evolved loon-like adaptations independently for diving and swimming. Their legs are similarly set far back on the body, with compressed, bladelike tarsi. The toes of grebes are not webbed, however, but are individually broadened and lobed. The plumage is dense and silky, and the tail is rudimentary. Many species have an elaborate and often spectacular form of aquatic courtship display. Although highly adapted for aquatic life, most grebes are strong fliers and northern species are migratory. There is one flightless species, *Centropelma micropteryum*, found at Lake Titicaca in the Andes. The family as a whole is virtually cosmopolitan. See AVES.

[K.C.P.]

## Podocopa

A suborder of the order Ostracoda in which the members have a shell that is without a permanent anterior aperture. The shells of different species vary in size, shape, and sculpture. Both pairs of antennae are used for locomotion, either swimming or crawling. The mandible is well developed and is usually provided with a palp of four podomeres. The palp also has a bunch of setae at its base. The heart is absent. A simple eye, or ocellus, is frequently present. Each maxilla is provided with a large respiratory plate. There are four pairs of postoral appendages: first, second, and third thoracic legs, and the furcae.

**Extant ostracods.** Four families of the Podocopa are recognized, based on the structure of the thoracic legs, and on the presence or absence and degree of development of the caudal ramus. These families are the Cypridae, Darwinulidae, Cytheridae, and Bairdiidae.

**Cypridae.** This family includes both marine and fresh-water genera in which all the thoracic legs are different. The first thoracic appendage is modified for mastication and the endopodite of this leg,

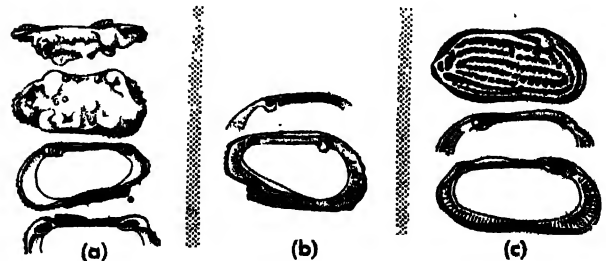
in the male, forms a prehensile palp. The third leg is bent dorsally and is used in keeping the body and inner surface of the shell free of foreign matter. The abdominal appendages and furcal rami, each have two claws and two setae, as a rule. In the Cypridopsinae, the rami are reduced to flagella-like setae. The Cypridae show definite seasonal cycles. Some species occur only in autumn and others in the spring or summer. Organisms that inhabit permanent bodies of water may have two or three generations a year; species living in temporary ponds and ditches generally have only a single generation a year. The life span varies with the species and with climatic conditions; ostracods that live in permanent bodies of water generally have longer life spans than those inhabiting temporary ponds and ditches.

The family Cypridae is divided into eight subfamilies which are widely distributed. Most genera and numerous species are Holarctic, or cosmopolitan.

**Darwinulidae.** This family has the first thoracic appendage modified for mastication, while the second and third thoracic legs are similar and adapted for crawling. The antennae are without swimming setae. The furcae are completely lacking, the body terminating in a single cone-shaped projection. The female is viviparous, that is, bears living young. A single genus, *Darwinula*, is known, with two species presently recognized. Only *Darwinula stevensoni* Brady and Norman, which inhabits the bottom of large fresh-water lakes, has been observed in North America.

**Cytheridae.** These are principally marine forms. The antennae of members of this family are without natatory setae and the three pairs of thoracic legs are similar, being adapted for crawling. The furcal rami are reduced and the ductus ejaculatorius is absent. Some species are viviparous. Species of the genus *Entocythere* live as commensals on the gills of various species of crayfish.

**Bairdiidae (Nesideidae).** These are entirely marine and possess a shell of firm consistency. The valves are conspicuously unequal, with the left being the larger. The antennae are well developed, but are not adapted for swimming. The three tho-



(a) *Oligocythereis fullonica* (Jones and Sherborn), a Jurassic cytheracean ostracod with entomodont hinge. (b) *Alatacythere alata* (Bosquet), an Eocene cytheracean ostracod with lobodont hinge. (c) *Legumncythereis corrugata* Le Roy, a Pliocene cytheracean ostracod with amphidont hinge.

Location	Type of hinge					
	Lophodont	Merodont	Entomodont	Lobodont	Schizodont	Amphidont
<b>Left valve</b>						
Anterior	Groove	Loculate groove	Loculate groove	Loculate pit	Biloculate socket	Socket
Anteromedian	Ridge	Denticulate ridge	Short dentate ridge	Lobate boss	Bifid tooth	Conical tooth
Median	Ridge	Denticulate ridge	Long denticulate ridge	Smooth or denticulate bar	Denticulate bar	Smooth or denticulate bar
Posterior	Groove	Loculate groove	Loculate groove	Loculate groove	Loculate socket	Loculate socket
<b>Right valve</b>						
Anterior	Ridge	Dentate ridge	Dentate ridge	Lobate boss	Bifid, stirpate tooth	Stirpate tooth
Anteromedian	Groove	Locellate groove	Short, wide loculate groove	Loculate pit	Biloculate socket	Socket
Median	Groove	Locellate groove	Long, narrow locellate groove	Smooth or locellate groove	Locellate groove	Smooth or locellate groove
Posterior	Ridge	Dentate ridge	Dentate ridge	Dentate ridge	Lobate reniform tooth	Lobate reniform tooth

\* After P. C. Sylvester-Bradley

racic appendages are similar in structure and all are adapted for locomotion. The caudal rami are very small and extremely mobile. The ejaculatory ducts are lacking. *Bairdia*, *Nesidea*, and *Bythocypris* are the principal genera in this family. Species have been collected from marine habitats in widely separated geographical regions. [L.F.]

**Extinct ostracods.** The Podocopa (Ordovician Recent) have many fossil representatives, both marine and fresh-water. The superfamily Cytheracea, in particular, is abundant in the Mesozoic and Cenozoic, and has many taxonomic divisions. Nearly half of the fossil species described belong to this taxon.

Hingement, which is important in their classification, may be either simple, with a groove in one valve accommodating the edge of the other, or compound, with the hinge of each valve divided into three or four elements. Six types of compound hinge are distinguished: (1) lophodont, (2) merodont, (3) entomodont, (4) lobodont, (5) schizodont, and (6) amphidont. This series is more or less morphogenetic, and certain evolutionary lines following it have been discovered. In general, the anteromedian element developed secondarily, and the terminal elements became dentate and then fused, changing from a ridge-and-groove to a tooth-and-socket arrangement. [R.V.K.]

**Bibliography:** E. Ferguson, Studies of the seasonal life history of three species of fresh-water Ostracoda, *Am. Midland Naturalist*, 32(3):713-727, 1944; R. W. Pennak, *Fresh-water Invertebrates of the United States*, 1953; G. O. Sars, *An Account of the Crustacea of Norway*, vol. 9, 1928.

## Poecilosclerida

An order of sponges of the class Demospongiae in which the skeleton includes two or more types of megascleres, each localized in a particular part of the sponge colony. Frequently one type of megasclere is restricted to the dermis; another type occurs in the interior of the sponge. Sometimes one category is embedded in spongin fibers; a second category, usually spinose, protrudes from the fibers at right angles. Spongin is always present but varies in amount from species to species. Microscleres are usually present; often several types occur in one species. A wide variety of microsclere categories is found in the order, but asters are never present.

In shape, poecilosclerid sponges are encrusting massive, lobate, or branching. Deep-sea species of ten have bizarre shapes. Sponges of this order are

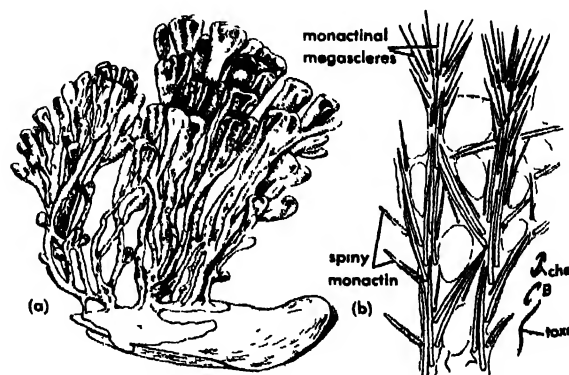


Fig. 1. (a) *Microciona prolifera*. (b) Spiculation of the same sponge. (After Hyman, 1940)

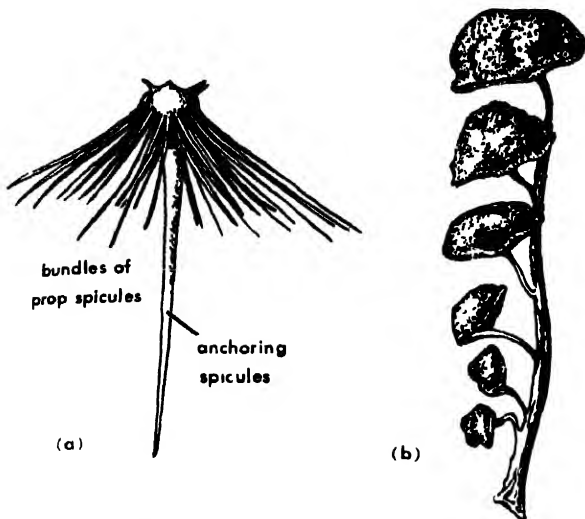


Fig. 2. Deep-sea Poecilosclerida. (a) *Cladorhiza longipinna*, 5500 meters. (b) *Esperipsis challengerii*, 1500 meters (from Hyman, 1940). (After Ridley and Dendy, 1887)

found in all seas and range from tidal waters to depths of at least 5500 meters.

Fossil sponges with skeletons comparable to those of Recent poecilosclerids are scattered through the fossil record from Cambrian times. Undoubted poecilosclerid spicules are known from tertiary deposits. See DEMOSPONGIAE. [W.D.H.]

### Poeobioidae

A phylum proposed by W. Fisher in 1946 for a single species *Poeobius meseres* Heath. A critical study of this particular species revealed it to be an aberrant polychaete. [C.B.C.]

**Bibliography:** H. Heath, A connecting link between the Annelida and the Echiuroidea (*Gephyrea armata*), *J. Morphol.*, 49:223-249, 1930; G. E. Pickford, Histological and histochemical observations upon an aberrant annelid, *Poeobius meseres* Heath, *J. Morphol.*, 80:287-319, 1947.

### Pogonophora

A group of animals regarded as the single class of the phylum Brachiata. They are characterized by the following features. The elongate body consists of three segments. The two anterior segments are fused to form an anterior region. This is separated by a diaphragm from the long trunk region regarded as a third segment. The trunk is subdivided into a preannular section consisting of a short metameric portion followed by a long nonmetameric portion, and a postannular section which has been considered to be secondarily metameric. Each segment has a separate coelom or body cavity and the first segment has coelomoducts which may represent an excretory organ. The coelom of the second segment is without ducts, but the third has well developed coelomoducts which serve as gonoducts for the single pair of gonads. There is no mouth, anus, or digestive canal. Some species may exceed 6

inches in length. There are from one to 223 tentacles on the first segment. Setae may occur in groups; these are quite different from annelid setae.

A ganglionic mass of nerve cells is situated in a cephalic lobe in the first segment and a longitudinal nerve cord, in the epidermis, lies on the side regarded as dorsal.

There are longitudinal dorsal and ventral blood vessels and a ventral heart. Blood is said to flow forward in the ventral vessel and back in the dorsal.

The sexes are separate. The testes are enormous and, with their ducts, fill the entire posterior half of the trunk. In females, the ovaries occupy the anterior half of the trunk. The eggs are large, few in number, elongate, bilaterally symmetrical, and heavily yolked. Cleavage is total, unequal, and determinate. It is neither spiral nor radial but unique to this group. There is no blastopore but the animal pole becomes anterior. The coelom is enterocoelic and the three segments become distinct early. There are no free-swimming larvae in the species described.

The Pogonophora have been obtained only from the sea bottom and usually at great depths, beginning with the discovery of *Lamellisabella* by P. Uschakov. In 1933 he found this species in the Sea of Okhotsk at a depth of 3500 meters. Many specimens have since been taken in many seas, both arctic and tropical. They occupy lamellated, chitin-like, nonbranching tubes and are apparently sedentary. The mode of feeding is highly conjectural because of the absence of all internal alimentary structures usually found in higher animals that are nonparasitic. It has been proposed that the tentacles catch fine, particulate matter which is digested extracellularly.

Two orders are recognized, the Athecanephria and the Thecanephria, which contain five families and eight genera. Characters of taxonomic value are the adhesive platelets on the skin, the bridle, and the tentacles. In different genera, the base of the tentacular crown is either circular, horseshoe shaped, or spiral.

At first considered to be polychaetes and set apart as a class by K. Johansson in 1937, the Pogonophora are now generally treated as a phylum, the Brachiata. They are closely related to the Enteropneusta and Echinoderm-chordate complex (Deuterostomia). See ANIMAL KINGDOM; ATHECANEPHRIA; BRACHIATA; CLEAVAGE, EMBRYONIC; COELOM; ENTEROPNEUSTA; OVUM; THECANEPHRIA.

[T.H.B.]

### Poinsett's method

A method of describing, by means of geometrical construction, the motion of a rigid body with a point fixed in space and with zero torque or moment acting on the body about the fixed point. If a rigid body is constrained to rotate about a smooth fixed axis, under no moments except those due to the axis reactions, the motion is simply one of con-



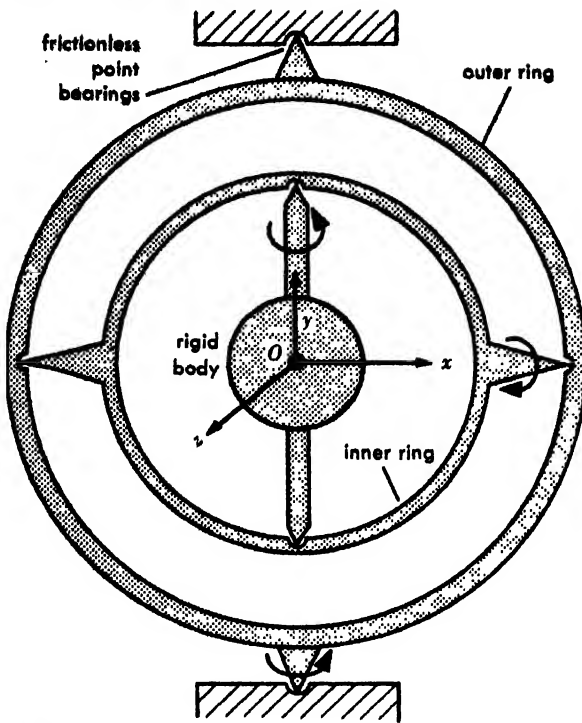


Fig. 1. Cardan's suspension. Point  $O$  of the rigid body is fixed in space while the body is free to rotate about any axis in space under no external moments.

stant angular velocity. If, however, the body is constrained to move with only one point fixed in space, the motion, even with no moment acting about that point, is much more complicated. Furthermore, the motion in this latter case is identical to that of a rigid body relative to its own center of mass, and it is, therefore, a more general type of motion for those cases where zero or negligible moments act about the center of mass. Such a body might be a top spinning on a frictionless table in a gravityless system, a body mounted within a Cardan suspension, or a spinning rocket flying in space outside of the atmosphere but in a gravity field. See CENTER OF MASS.

**Cardan's suspension.** Consider a heavy body mounted with only one point fixed, constructed from light rings in an arrangement known as Cardan's suspension. In Fig. 1, point  $O$  is the fixed point, and it is assumed that the frictional torques can be made negligible and that the mass of the suspension system compared with the heavy body is negligible.

Let  $O$  be the center of a coordinate system composed of the principal axes of the body  $x, y, z$ , with unit directional vectors  $i, j, k$ , respectively.

The vector angular velocity  $\omega$  and vector angular momentum  $H$  of the body are

$$\omega = \omega_x i + \omega_y j + \omega_z k \quad H = I_x \omega_x i + I_y \omega_y j + I_z \omega_z k \quad (1)$$

where  $I_x, I_y, I_z$  are moments of inertia about the  $x, y, z$  axes respectively, and the products of inertia are zero about the principal axis (see RIGID-BODY

DYNAMICS). The kinetic energy of the body is constant and is

$$T = \frac{I_x \omega_x^2 + I_y \omega_y^2 + I_z \omega_z^2}{2} \quad (2)$$

The angular momentum  $H$  is constant in magnitude and direction. Its magnitude is given by

$$H^2 = (I_x \omega_x)^2 + (I_y \omega_y)^2 + (I_z \omega_z)^2 \quad (3)$$

From (1) and (2),

$$\omega \cdot H = 2T \quad (4)$$

**Ellipsoid equations.** If now a line  $OA$ , called the invariable line, is drawn in the fixed direction of  $H$  (Fig. 2), and  $OB$  is the vector angular velocity  $\omega$  at any instant, then the line  $BC$ , drawn perpendicular to  $OA$ , determines line  $OC$  such that

$$OC = \frac{\omega \cdot H}{H} \quad (5)$$

From (4) and (5),

$$OC = \frac{2T}{H}$$

Therefore,  $C$  is a fixed point and the plane through  $C$  normal to  $OA$  is a fixed plane, called the invariable plane. The terminus of  $\omega$  (point  $B$ , Fig. 2) moves on the invariable plane during motion of the rigid body.

If point  $B$  is given coordinates  $x_1, y_1, z_1$ , then

$$\omega = ix + jy + kz$$

and Eq. (2) becomes

$$I_x x^2 + I_y y^2 + I_z z^2 = 2T \quad (6)$$

Equation (6) is the equation of the Poinsot ellipsoid which is fixed in the rigid body tangent to the invariable plane at  $B$ , and with center at  $O$ . The semiaxes are given by

$$a = \sqrt{\frac{2T}{I_x}} \quad b = \sqrt{\frac{2T}{I_y}} \quad c = \sqrt{\frac{2T}{I_z}}$$

As the body moves, the ellipsoid rolls on the invariable plane.

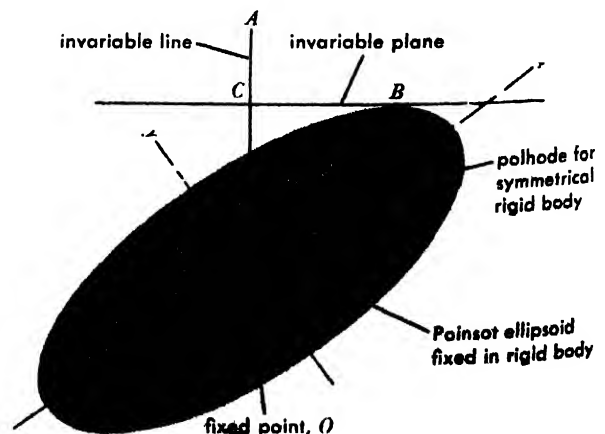


Fig. 2. Poinsot ellipsoid.

able plane because it has an angular velocity vector which terminates at  $B$ , the point of contact of the ellipsoid and plane (see Fig. 2).

The point  $B$  is called the pole of the axis of rotation. The locus of this pole on the ellipsoid was called by L. Poinso the polhode and the locus on the tangent plane the herpolhode.

The resulting motion is, in general, quite complicated except when the Poinso ellipsoid is a surface of revolution. If the invariable plane were represented by a sheet of paper and if an ellipsoid given by Eq. (6) were constructed, coated with ink, and rolled on the paper holding center  $O$  fixed, the ink would trace out a herpolhode curve. On joining the fixed point  $O$  to all points on this curve, the space cone of the  $\omega$  vector would result. In this way the motion of the rigid body under no moments can be visualized and represented.

[R.E.BO.]

*Bibliography:* H. Goldstein, *Classical Mechanics*, 1950.

## Point

In axiomatic geometry, usually a completely undefined (primitive) element, although there are axiomatizations of geometry in which those properties of "point" that are desired are given by postulates. To illustrate: in a geometry in which line is a primitive element, points may be defined as classes of lines that conform to certain requirements (postulates). In  $n$ -dimensional metric analytic point geometry, a point may be defined as an ordered  $n$ -tuple of numbers. In analytic projective point geometry of  $n$  dimensions, the introduction of the so-called infinite or ideal points is accomplished by defining a point as an ordered  $(n+1)$ -tuple of numbers, not all zero, and identifying points  $(x_1, x_2, \dots, x_n, x_{n+1})$ ,  $(y_1, y_2, \dots, y_n, y_{n+1})$  if and only if  $x_i = \rho \cdot y_i$ , with  $i = 1, 2, \dots, n+1$  and  $\rho \neq 0$ . The ideal points are those  $(n+1)$ -tuples of numbers with the  $(n+1)$ st one zero. See ANALYTIC GEOMETRY; GEOMETRY, EUCLIDEAN. [L.M.BL.]

## Point source

In discussing radiation, it is convenient to define a concept called the point source, that is, a source having definite position but no extension in space. If the radiation propagates in radially straight lines (or, which is the same thing, in spherical waves) from the point source, conservation of energy demands that the intensity of the radiation decrease in any direction inversely as the square of the distance from the source. No physical source is actually a mathematical point, but for distances sufficiently large compared to the dimensions of the source, the inverse-square law may be a good approximation. See INVERSE-SQUARE LAW. [M.H.H.]

## Point-contact diode

A semiconductor rectifier using the barrier formed between a specially prepared semiconductor surface and a metal point to produce the rectifying ac-

tion. The contact is usually maintained by mechanical pressure, but in some instances it may be welded or bonded. The rectifying action implies that the resistance of the contact is significantly greater for one direction of applied voltage (reverse direction) than for the other (forward direction).

Point-contact diodes have been widely used in radio and television, and most notably in computers and in microwave detectors and ultra-high-frequency mixers.

Whenever a metal-semiconductor contact is made an electrical barrier generally exists between the two. Only specially prepared ohmic contacts show no barrier. This barrier impedes the flow of majority carriers. (See SEMICONDUCTOR for a definition of majority carriers and a discussion of conductivity type.) In an  $n$ -type semiconductor the majority electrons are immobilized by the barrier, and a bias, which renders the semiconductor positive with respect to the metal, repels the positive holes (electron vacancies) in the metal. Very little current flows under this condition and the resistance is high. If the  $n$ -type semiconductor is negative with respect to the metal, most of the electrons are still immobilized by the barrier, but holes can now enter from the metal and a relatively large current flows, that is, low resistance is present. For a  $p$ -type semiconductor the barrier impedes holes, and the bias polarities are reversed for the high- and low-resistance conditions.

The small physical size of the point contact forces a high current density in the neighborhood of the point. The current distribution in the semiconductor gives rise to a spreading resistance in series with the barrier. The forward current is limited by this resistance while the reverse current is limited by the barrier. The spreading resistance steadily decreases with increasing forward current because of heating and the injection of minority carriers. See TRANSISTOR.

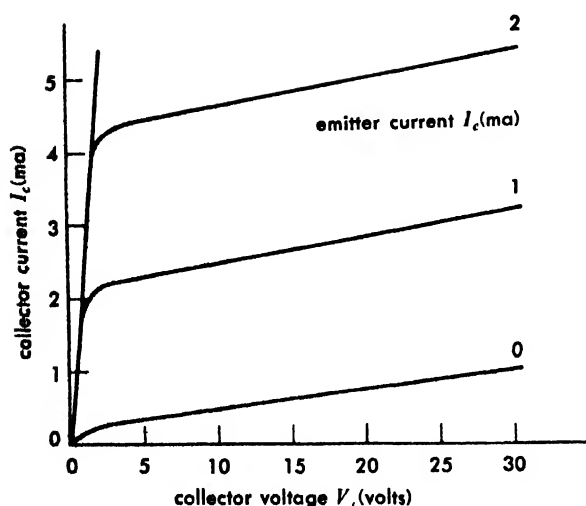
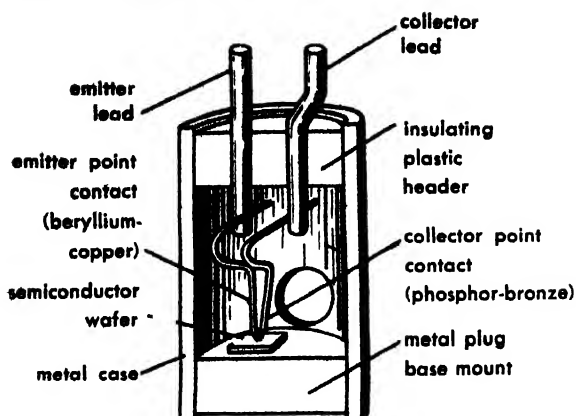
In the reverse direction there is a breakdown phenomenon at relatively high voltages due either to heating or avalanching of the minority carriers passing through the high field barrier region.

[L.P.HU.]

## Point-contact transistor

A transistor in which the emitter and collector consist of metal point contacts closely spaced on the surface of a block of semiconductor. The usual configuration is with both points on the same surface and about 2 mils apart (see illustration), although good devices have also been made with the points on opposite sides of a thin wafer of semiconductor.

This type of transistor was the first transistorlike device invented. The most common type uses  $n$ -type semiconductor material, beryllium-copper emitter-point material, and phosphor-bronze collector-point material. In fabrication the surface of the semiconductor is carefully lapped and etched. The sharp points are mechanically assembled with some spring



Point-contact transistor cutaway and characteristics.

pressure against the surface. The collector point is electrically pulsed in the reverse-bias direction with sufficient voltage and total energy to cause electrical breakdown. The point of contact is heated nearly to the melting point of the semiconductor. The pulse duration is a millisecond or less. The result of this electric forming procedure is to increase the current multiplication factor  $\alpha^*$  of the collector point from something much less than unity to the order of 10.0. The injection efficiency  $\gamma$  of the point emitter is about 0.3 and the transfer efficiency is about 1.0 so that the over-all current gain  $\alpha$  of the device is 3.0. See TRANSISTOR.

The electrical forming process, besides increasing  $\alpha^*$ , also increases the collector barrier leakage current  $I_{co}$  so that, at a collector voltage of 10 volts, a typical point-contact device will draw 1 milliamperes in the absence of emitter current. This compares to 1 microampere for a junction transistor under the same condition.

Point-contact transistors can be made with frequency ranges up to 100 megacycles and power ratings of 200 milliwatts. They can be used quite conveniently for oscillators and flip-flops, because their  $\alpha > 1.0$  causes them to show a negative-resist-

ance characteristic when the base is used as an input. They have not achieved widespread acceptance because of the variability of their characteristics and because of their relatively high  $I_{co}$ .

[L.P.HU.]

*Bibliography:* W. Shockley, *Electrons and Holes in Semiconductors*, 1950.

## Poison

A substance which by chemical action and at low dosage can kill or injure living organisms. Broadly defined, poisons include chemicals toxic for any living form: microbes, plants, or animals. For example, antibiotics like penicillin, although nontoxic for mammals, are poisons for bacteria. In common usage the word is limited to substances toxic for man and mammals, particularly where toxicity is a substance's major property of medical interest. Because of their diversity in origin, chemistry, and toxic action, poisons defy any simple classification. Almost all chemicals with recognized physiological effects are toxic at sufficient dosage. The same compound may be considered a drug or a poison, depending on dosage, effect, or intended use.

**Origin and chemistry.** Many poisons are of natural origin. Some bacteria secrete toxic proteins (for example, botulinus, diphtheria, and tetanus toxins), among the most poisonous compounds known (see TOXIN, BACTERIAL). Lower plants notorious for poisonous properties are ergot (*Claviceps purpurea*) and a variety of toxic mushrooms. Ergot, a fungal parasite of rye, has been the source of numerous epidemics of poisoning from the use of contaminated rye flour. The fungus contains many different alkaloids, some of which are also useful drugs. Among the best known toxic mushrooms are the fly agaric (*Amanita muscaria*) containing muscarine, and the destroying angel (*Amanita phalloides*) whose toxic agents are phalloidin and amatoxin. See MUSHROOM.

Higher plants, which constitute the major natural source of drugs, contain a great variety of poisonous substances. Many of the plant alkaloids double as drugs or poisons, depending on dose. These include curare, quinine, atropine, mescaline, morphine, nicotine, cocaine, picrotoxin, strychnine, lysergic acid, and many others. Some of the alkaloids were used in classical antiquity (coniine was the toxic agent of the extract of spotted hemlock, *Conium maculatum*, drunk by Socrates). Some are of prehistoric antiquity (quinine and curare were used by South American Indians before the advent of Europeans) and some date from our earliest records of man (the opium poppy is believed to have been cultivated in the Stone Age). See ATROPINE; COCAINE; MORPHINE; QUININE.

Poisons of animal origin (venoms) are similarly diverse. Toxic marine animals alone include examples of every phylum from Protozoa (dinoflagellates) to Chordata (a number of fishes). Insects and snakes represent the best known venomous land animals, but on land all phyla include

poison-producing species. Among mammalian examples are certain shrews with poison-producing salivary glands. See DINOFLAGELLIDA.

Poisons of nonliving origin vary in chemical complexity from the toxic elements, for example, the heavy metals, to complex synthetic organic molecules. Most of the heavy metals (gold, silver, mercury, arsenic, lead) are poisons of high potency in the form of their soluble salts. Strong acids or bases are toxic largely because of corrosive local tissue injury; ingestion, for example, sometimes results in fatal perforation of the gastrointestinal tract. Many other elemental substances are toxic at low concentrations: selenium, beryllium, cadmium, manganese, phosphorus, and zinc. In general, it may be anticipated that any chemically reactive element is likely to be toxic to man unless it represents one of the bulk elements in the body, in the ionic form in which it exists in the body.  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Ca}^{++}$ , and  $\text{Cl}^-$ , for example, are relatively nontoxic.

The chemically reactive gases, hydrogen sulfide, hydrocyanic acid, chlorine, bromine, and ammonia are also toxic, even at low concentration, both because of their corrosiveness and because of more subtle chemical interaction with enzymes or other cell constituents. An example is the toxic action of cyanide ion, probably explained by its inactivation of cytochrome oxidase. See CYTOCHROME.

Many organic substances of synthetic origin are highly toxic and represent a major source of industrial hazard. Most organic solvents are more or less toxic on ingestion or inhalation. Ethanol is a relatively innocuous exception, in part because available enzymatic reactions channel it into major normal metabolic pathways along which it can be oxidized to products familiar to mammalian metabolism, such as acetate. Many other alcohols, such as methanol, are much more toxic. Many solvents (for example, carbon tetrachloride, tetrachloroethane, dioxane, and ethylene glycol), produce severe chemical injury to the liver and other viscera, sometimes from rather low dosage. With certain poisons of this group, a high degree of variation in individual susceptibility exists.

**Chemical correlations.** Since poisons represent all chemical classes from the elements to complex alkaloids and large proteins, general chemical constitution has no defining value for toxicity. However, in limited instances chemical features correlate with toxic action. Many of the chlorinated hydrocarbons, for example (carbon tetrachloride) have similar toxicity for liver, heart and kidneys. A number of alkyl phosphates (diisopropylfluorophosphate, tetraethylpyrophosphate and related compounds) are very potent inhibitors of the enzyme acetylcholinesterase and produce a consistent set of physiological changes arising from the inactivation of this enzyme. A number of tertiary amino esters of aromatic acids (procaine, cocaine) are local anesthetics and also share marked cardiac and nervous system toxicity. The digitalis glyco-

sides, all represented by steroids condensed with sugars and containing an additional lactone ring, share many pharmacological properties, notably characteristic effects on cardiac function which make these drugs cardiotoxins at low concentration.

**Physiological actions.** The action of poisons is generally described by the physiological or biochemical changes which they produce. For most poisons, a descriptive account can be given which indicates what organic system (for example, heart, kidney, liver, brain, bone-marrow) appears to be most critically involved and contributes most to seriously disordered body function or death. In many cases, however, organ effects are multiple, or functional derangements so generalized that a cause of death cannot be localized.

Although a comprehensive list cannot be presented here, some illustrations may be given. Anoxia, lack of oxygen for cellular respiration, can result from poisons acting at different sites. Phosphene or other inhaled corrosive vapors may produce massive flooding of the lungs with edema fluid and thus cause mechanical suffocation. Carbon monoxide acts by binding to hemoglobin with a much greater affinity than oxygen and so produces anoxia by interference with oxygen transport. Cyanide prevents tissue respiration at a terminal intracellular site by poisoning cytochrome oxidase which catalyzes the final step of oxygen utilization. Central nervous system depression, with coma and ultimate respiratory and circulatory failure, results from a large group of drugs: the general anesthetics (ether, chloroform, cyclopropane, ethylene), the barbiturates, the opium alkaloids (morphine, codeine, and related compounds), and less regularly from many other poisons. Liver injury with jaundice (toxic hepatitis) is a prominent result of many poisons, characteristically the chlorinated hydrocarbons, elemental phosphorus, certain heavy metals (arsenic, antimony, barium, copper) and a large diversity of organic compounds. Kidney damage results from many toxic chemicals, but is common after exposure to carbon tetrachloride, mercuric salts, and ethylene glycol. Convulsions, sometimes fatal, are a common response to acute poisoning by diverse agents, but are a particularly characteristic effect of strychnine, atropine, ergot, alkyl phosphates, picrotoxin, cocaine, and certain snake venoms. Among the medically important toxic agents are those that reduce or exaggerate effective transmission of the nerve impulse across synaptic junctions, especially in the autonomic nervous system. All such compounds, sometimes called autonomic agents, act so as to mimic or to block the physiological effects of either of the two known neurohumoral transmitter substances of the motor branches of autonomic nerves, acetylcholine and norepinephrine. See ACETYLCHOLINE.

Poisons that act to exaggerate those physiological effects normally produced by acetylcholine release include muscarine and pilocarpine. The anti-

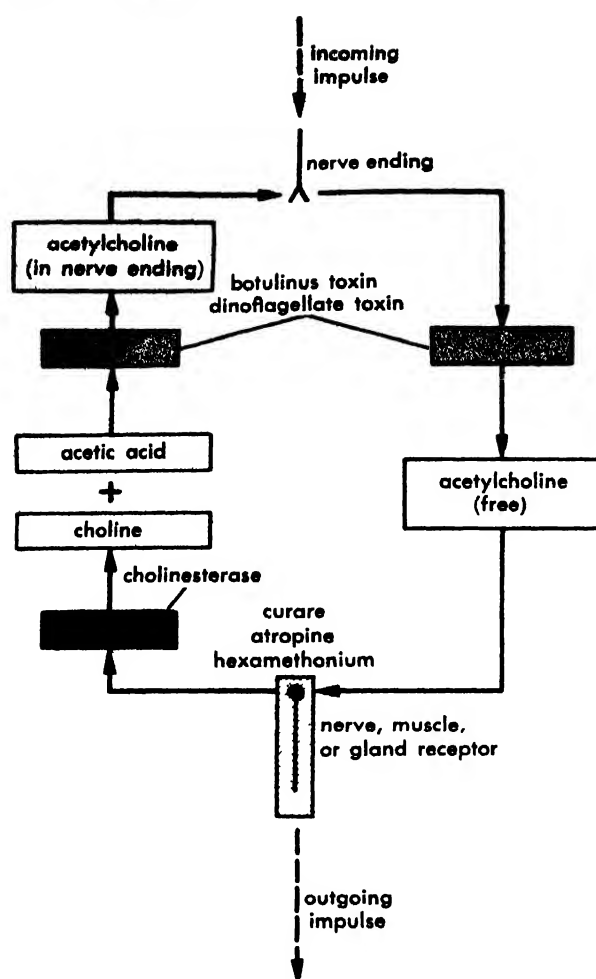


Fig. 1. Acetylcholine cycle, the target of many poisons. The "anticholinesterase" poisons prevent the breakdown of acetylcholine by inactivating cholinesterase (black). Botulinus and dinoflagellate toxins (gray) hinder the synthesis or the release of acetylcholine. Curare and atropine make the receptor (hatched gray) less sensitive to the chemical stimulus. (Adapted from *Scientific American*, 1959)

cholinesterases (neostigmine, physostigmine, and the alkyl phosphates) have the same general physiological effects, although by a different mechanism. These compounds slow the heart, lower blood pressure, increase secretion of fluids into the respiratory tract, and narrow the respiratory passages. Toxic doses may produce death by suffocation or by failure of the circulation. Poisons that prevent acetylcholine from producing skeletal muscle contraction are curare and a number of related synthetic compounds used in anesthesia (succinylcholine, decamethonium). These drugs result in weakness of all voluntary muscle and may cause death through paralysis of respiratory muscles. Exaggeration of the normal effects of norepinephrine and epinephrine (adrenergic drugs) may produce dangerously high blood pressure, rapid heart action, and sometimes fatal disturbances of cardiac rhythm. See EPINEPHRINE.

Many normal functions (glandular secretion and contraction of voluntary and smooth muscle) depend on the cyclic release, breakdown, and resynthesis of acetylcholine at the endings of cholinergic (acetylcholine secreting) nerves. The accompanying illustration indicates the different ways in which several distinct poisons may interfere with the normal operation of the acetylcholine cycle. In the acetylcholine cycle, an impulse reaching a nerve ending liberates acetylcholine which stimulates a receptor. The receptor is freed for further impulses by the enzyme cholinesterase, which breaks down acetylcholine into acetic acid and choline. These are resynthesized by other enzymes into new acetylcholine.

**Mechanism of action.** More precise understanding of the mechanism of poisons requires detailed knowledge of their action in chemical terms. Information of this kind is available for only a few compounds, and then in only fragmentary detail. Poisons that inhibit acetylcholinesterase have toxic actions traceable to a single blocked enzyme reaction, hydrolysis of normally secreted acetylcholine. Detailed understanding of the mechanism of chemical inhibition of cholinesterase is not complete, but allows some prediction of chemical structures likely to act as inhibitors.

Carbon monoxide toxicity is also partly understood in chemical terms, since formation of carboxyhemoglobin, a form incapable of oxygen transport, is sufficient to explain the anoxic features of toxicity.

Heavy metal poisoning in many cases is thought to involve inhibition of enzymes by formation of metal mercaptides with enzyme sulphydryl groups, the unsubstituted form of which is necessary for enzyme action. This is a general reaction that may occur with a variety of sulphydryl-containing enzymes in the body. Specific susceptible enzymes whose inhibition explains toxicity have not yet been well documented.

Metabolic antagonists active as poisons function by competitive blocking of normal metabolic reactions (see ENZYME INHIBITION). Some antagonists may act directly as enzyme inhibitors, others may be enzymatically altered to form derivatives which are even more potent inhibitors at a later metabolic step. An example of the latter is the biosynthetic incorporation of metabolite analogs into much more complex molecules, particularly the incorporation into nucleic acids of altered purine or pyrimidine bases such as 8-azaguanine and 5-bromouracil, or the incorporation into proteins of altered amino acids such as *p*-fluorophenylalanine or 7-azatryptophan.

A well-studied example of biosynthetic production of a highly toxic metabolic analog arose from studies of fluoroacetate, a toxic component of plants of the genus *Dichapetalum*. Fluoroacetate is enzymatically converted to fluorocitrate, through reactions analogous to those normal steps leading from acetate to citrate. The fluorocitrate formed




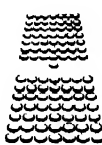
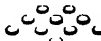


toxicity rating	practically nontoxic	slightly toxic	moderately toxic	very toxic	extremely toxic	supertoxic	supremely toxic
examples	glycerin, water, graphite, lanolin	ethyl alcohol, lysol, castor oil, soaps	methyl (wood) alcohol, kerosene, ether	tobacco, aspirin, boric acid, phenol, carbon tetrachloride	morphine, bichloride of mercury	potassium cyanide, heroin, atropine	botulinus toxin, some snake venoms
probable lethal dose, mg/kg	more than 15,000	5000 to 15,000	500 to 5000	50 to 500	5 to 50	less than 5	less than 0.5
probable lethal dose for a 70 kg (155 lb) man	 more than 1 qt	 1 pt-1 qt	 1 oz-1 pt	 1 tsp-1 oz	 7 drops-1 tsp	 a taste (less than seven drops)	 a taste (less than seven drops)

Fig 2. Toxicity-rate scale of substances according to the size of the probable lethal dose. Many drugs fall into the highly toxic categories. Drawings suggest the fatal doses of water, whiskey, ether, aspirin, morphine,

and cyanide, the last three depicted in terms of aspirin-sized tablets. The scale was suggested by Marion N. Gleason, Robert E. Gosselin, and Harold C. Hodge. (Adapted from *Scientific American*, 1959)

from fluoroacetate is a potent inhibitor of aconitase, a citrate-utilizing enzyme in the tricarboxylic acid cycle, and this inhibition can explain the toxicity of fluoroacetate.

Where poison mechanisms are relatively well understood it has sometimes been possible to employ rationally selected antidotes. Thus the clarification of arsenic as a sulfhydryl poison led to the introduction of the effective metal-poisoning antidote, dimercaptopropanol (British antilewisite). Knowledge of the physiology of anticholinesterase poisons has permitted the use of rational physiological antagonists of excessive cholinergic stimulation, such as atropine and curare. On a more chemical basis, recent studies of the enzymatic mechanism of cholinesterase have led to more direct antidotes (pyridine aldoxime methiodide and other oximes) for the anticholinesterase poisons. These antidotes act by removing the inhibitor from the enzyme.

It should be noted that for the great majority of poisons the mechanism of toxic action is not as well understood as in the selected examples above. In only a few cases are poisons known to act directly by altering enzyme function. Other poisons, probably much more numerous, may involve reactions with less well-defined cellular components.

**Potency.** The strength or potency of poisons is most frequently measured by the lethal dose, potency being inversely proportional to lethal dose. Although killing dose represents a useful endpoint for potency appraisal, much smaller doses of a given poison may produce toxic changes in an organism. In estimating potency, a large test population should be used because of inherent individual variation in susceptibility. The relationship between per cent of a population killed and dose of poison sometimes, but not often, conforms to a sym-

metrical sigmoid curve representing the integrated or cumulative form of the bell-shaped Gaussian distribution curve. From statistically treated dose-response data, the dose killing 50% of the sample population can be determined, and is usually designated the MLD (median lethal dose) or  $LD_{50}$ . This is the commonest measure of toxic potency, although for special purposes the lethal dose for other fractions of the population may also be used (for example  $LD_{10}$ ,  $LD_{90}$ ). The spread of the distribution (for example dose difference between  $LD_1$  and  $LD_{99}$ ) is highly variable for different poisons and is of course also greatly influenced by the relative heterogeneity or homogeneity of the test population. See LETHAL DOSE 50.

**Health aspects.** The environment of civilized man abounds in toxic chemicals: drugs, industrial products, fuels, pesticides, and many common household products such as cleaning agents and paint products. Acute and chronic poisoning represent an appreciable public health hazard. In 1955 approximately 8000 lives were lost from exposure to harmful chemicals, while the estimated number of nonfatal accidents was at least 1,000,000. Of the fatal cases of acute poisoning in 1955, approximately half were suicidal and half accidental; homicidal poisoning represented less than 0.5% of the total.

The multiplication of toxic chemicals in everyday use led, in 1953, to the establishment of information centers known as Poison Control Centers, which provide physicians with emergency telephone information concerning toxicity and antidotes for common poisons as well as for brand-named chemical formulations. These services were greatly assisted by the publication in 1957 of a compilation by M. N. Gleason and coauthors of brand-named



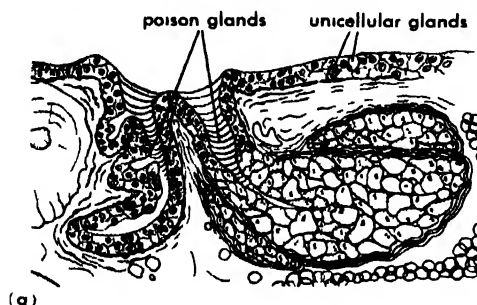
## Poison gland

products with information as to ingredients and quantitative toxicity. The authors of this book also proposed a useful scale of toxicity, here reproduced in modified form, with examples of specific toxic classes. See TOXICOLOGY. [E.AD.]

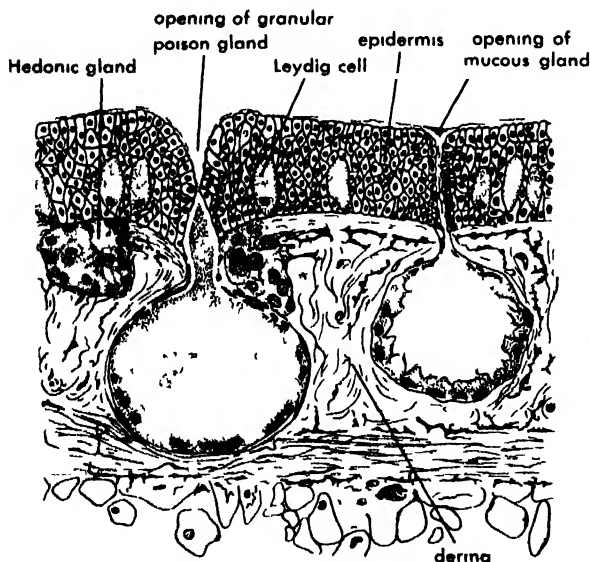
**Bibliography:** L. T. Fairhall, *Industrial Toxicology*, 2d ed., 1957; M. N. Gleason, R. E. Gosselin and H. C. Hodge, *Clinical Toxicology of Commercial Products*, 1957; W. S. Spector, *Handbook of Toxicology*, vol. 1, 1956; W. F. Von Oettingen, *Poisoning, A Guide to Clinical Diagnosis and Treatment*, 2d ed., 1958.

## Poison gland

The specialized gland of certain fishes (illustration a), as well as the granular and some mucous glands of many aquatic and terrestrial Amphibia



(a)



(b)

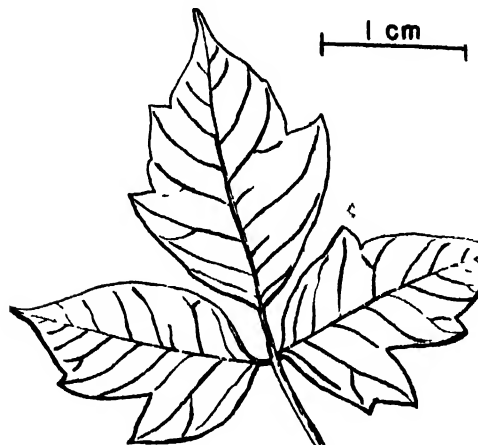
(a) Multicellular poison gland of catfish (modified from Reed). (b) Amphibian skin, generalized (modified from A. B. Dawson and G. K. Noble).

(illustration b). The poison glands of fishes are simple or slightly branched acinous structures which use the holocrine method of secreting a mucuslike substance. The poison glands of snakes are modified oral or salivary glands. Amphibian glands are simple, acinous, holocrine, with granular secretion. In some cases these amphibian poison glands produce mucus by a merocrine method

of secretion. These glands function as protective devices. See EPITHELIUM; GLAND. [O.E.N.]

## Poison ivy

A general name applied to certain species of the genus *Rhus*, in the sumac family (Anacardiaceae). *Rhus radicans* is the poison ivy of eastern North America; *Rhus diversiloba* is the poison oak of Cal



Leaf of poison ivy. (L. H. Bailey, *The Standard Cyclopaedia of Horticulture*, vol. 3, Macmillan, 1937)

ifornia. These plants are natives of North America. Both cause ivy poisoning, an annoying and often painful dermatitis.

*Rhus radicans*, the most wide-spread species, is extremely variable. It has a bushy or climbing habit. 3-foliolate leaves which are smooth and glossy or hairy, entire, toothed or lobed. Poison ivy bears white fruit which differs from the nonpoisonous sumacs with their red fruits. See HYPERSENSITIVITY, SAPINDALES. [P.D.S.]

## Poison sumac

This plant, *Rhus vernix*, is a member of the sumac family (Anacardiaceae). It is an inhabitant of swamps ranging from Quebec to Minnesota, and southward to Florida, Louisiana, and Texas. It is a tall bush or small tree bearing pinnately compound leaves with 7-13 entire (without marginal teeth) leaflets, and drooping, axillary clusters of persisting, white fruits. Like poison ivy, this plant is poisonous to touch, causing in many persons a severe



Poison sumac, *Rhus vernix*. (L. H. Bailey, *The Standard Cyclopaedia of Horticulture*, vol. 3, Macmillan, 1937)

inflammation of the skin, or dermatitis. The presence of white fruits separates this species from the nonpoisonous sumacs with their red fruits. See HYPERSENSITIVITY; SAPINDALES. [P.D.S.]

### Poisonous plants

Almost 400 species of vascular plants, representing 70 families, have been listed as poisonous plants occurring in the United States. About 100 of these species cause dermatitis. Among these are such common plants as poison ivy, *Rhus radicans*; poison sumac, *Rhus vernix*; and vipers bugloss, *Echium vulgare* (see POISON IVY; POISON SUMAC). Plants representing nearly a score of species possess spines, thorns, or fruits having sharp awns, which often cause injury to animals. See FRUIT (BOTANY); STEM (BOTANY). In addition to these are many other species more appropriately catalogued as poisonous plants because they contain toxic principles which initiate pathological conditions in animals.

Sixteen species are known to cause hydrocyanic poisoning. In this group are species of the genera *Sorghum* and *Prunus*. Wild cherry, *Prunus serotina*, is the most poisonous, and wilted leaves of

this plant contain the greatest amount of the toxic glucoside, amygdalin, which, when broken down by enzyme action, produces the lethal hydrocyanic acid. Symptoms of poisoning are staggering, convulsions, difficult breathing, bloating, and subsequent death, usually within an hour after eating the leaves.

In western North America, the two genera *Astragalus* and *Oxytropis* contain about 100 species, many of which are poisonous. In reference to the crazed reactions of the poisoned animals, these plants are called "loco weeds," the word loco, of Spanish origin, meaning crazy. The animal loses muscular control, appetite, flesh, and ultimately dies. Death camas, any one of a half-dozen species of the genus *Zygadenus*, is another lethal stock-poisoning plant of western North America. The toxic principle, consisting of one or more alkaloids, causes salivation, nausea, vomiting, staggering, difficult breathing, and sometimes a state of coma which may persist for days, followed by death.

Water hemlock, *Cicuta maculata*, and several other species of this genus contain a resinlike substance, cicutoxin, found chiefly in the roots or root-



poison hemlock



poison ivy



water hemlock



loco weed



pokeweed



black nightshade



Jimson weed



poison sumac



mayapple



sheep laurel



monkshood



death camas



belladonna



henbane

Poisonous plants. (Adapted from material in Webster's New International Dictionary, Second Edition, copy-

right 1934, 1939, 1945, 1950, 1953, 1954, 1957, 1959 by G. and C. Merriam Co.)

stocks, which is a lethal poison. See ROOT (BOTANY). The symptoms in man and animals are similar: nausea, vomiting, difficult breathing, and violent convulsions, ending in death caused by respiratory failure. Poison hemlock, *Conium maculatum*, contains in its fruits and leaves a volatile alkaloid, coniine, which causes a gradual lessening of muscular power, a rapid feeble pulse, and gradual lung paralysis resulting in death. See LEAF (BOTANY). Pokeweed, *Phytolacca americana*, has a poisonous root containing a toxic alkaloid and a glucoside which resembles a saponin. The seeds apparently contain these same poisons. See SEED (BOTANY). Symptoms of poisoning are vomiting, purging, spasms, and sometimes convulsions which lead to ultimate death caused by paralysis of the respiratory organs.

Plants containing the toxic alkaloids hyoscyamine and hyoscine are henbane, *Hyoscyamus niger*; belladonna, *Atropa belladonna*; jimson weed, *Datura stramonium*; and matrimony vine, *Lycium halimifolium*. Symptoms of poisoning are headache, dizziness, nausea, great thirst, failure of vision, and in the worst cases, mania, convulsions, and death. In this same family are several species of the genus *Solanum*, including European bitter-sweet, *S. dulcomara*, and black nightshade, *S. nigrum*, which contain the alkaloidal glucoside solanine. Solanine poisoning, in its commonest form, produces symptoms of narcosis and paralysis. In gastric poisoning, it causes salivation, vomiting, bloating, and diarrhea.

Certain species of monkshood, *Aconitum*, contain the alkaloids aconitine and aconine in all parts of the plant but particularly in the roots and seeds. These toxic principles cause muscular weakness, difficult breathing, weak pulse, and bloating. Poisoned horses and sheep often recover. A number of species of larkspur, *Delphinium*, contain several toxic alkaloids which cause animals to lose appetite, to stagger, and in severe cases, to fall and lie with legs rigidly extended. Mayapple, *Podophyllum peltatum*, contains a bitter, resinous, toxic substance, podophyllin, but because of the bitter taste, these plants are seldom eaten in amounts that are harmful to stock.

Sheep laurel, *Kalmia angustifolia*, and other species of this genus contain the toxic principle andromedotoxin which poisons sheep, goats, horses, and cattle. It causes salivation, increased nasal secretions, emesis with convulsions, and subsequently paralysis of the limbs. Animals may remain ill for 2 or more days and then recover.

Oleander, *Nerium oleander*, introduced in the southern United States as an ornamental shrub, contains two glucosides having properties similar to those of digitalis glucosides. Sheep, goats, horses, cattle, and poultry have been poisoned by eating the leaves. Symptoms of poisoning are nausea, vomiting, colic, vertigo, drowsiness, slowing of pulse, irregular heart action, bloody diarrhea, unconsciousness, respiratory paralysis, and finally death.

Various grasses including rye, barley, and occasionally wheat sometimes become infected with a fungus, *Claviceps purpurea* (see FUNGI). In the heads of such grasses, whole grains are replaced by dark, hard, cylindrical bodies called ergots, which eventually function in reproduction of the fungus. Ergots are poisonous to stock and if permitted to eat infected plants, either in pasture or in hay, the animals develop a diseased condition, becoming emaciated and often covered with sores. In the worst cases, portions of their tails or ears may be sloughed off. In females, abortion may occur. [P.D.S.]

Bibliography: W. C. Muenscher, *Poisonous Plants of the United States*, 2d ed., 1951.

## Polar meteorology

The application of meteorological principles to a study of atmospheric conditions in the earth's high latitudes or polar cap regions, northern and southern. These zones develop distinctive atmospheric character basically because of the obliquity of sun rays and the alternation of long periods of darkness and daylight (see GEOGRAPHY, MATHEMATICAL). Although solar radiation or its absence gives to the polar atmosphere its strongly contrasting winters and summers, other phenomena exert important influences. These include elevation, the nature of the earth's surface (soil, water, thin snow, thick ice), the size of continents and oceans, and their location with respect to circulation patterns.

**Winter.** During dark low-sun or winter seasons the absence of the sun permits cooling of the snow surface and overlying atmosphere. The earth's surface, because it is essentially a gray- or black-body radiator, cools more rapidly than the atmosphere. This causes the characteristic polar temperature inversion or temperature increase with height. In a sunless, cloudless atmosphere the extent of temperature inversion depends on the relative magnitude of compensating nonradiative heat fluxes, principally turbulent transfer of heat in the air and heat conduction in the snow.

**Surface temperature.** Rapid changes as well as extremes are characteristic. At the South Pole on

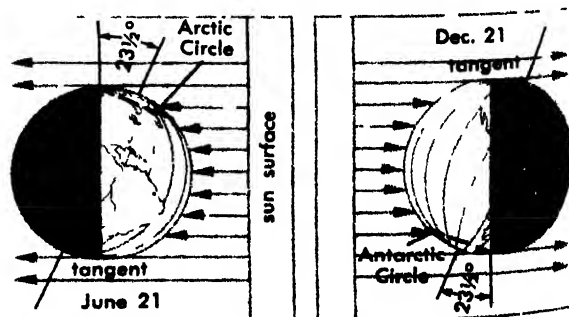


Fig. 1. The angular relationship between the direction of the sun's energy and the earth's surface changes during the year. Note the reversal of winter's predominant darkness, as against the summer's long hours of daylight, in northern and southern polar regions.

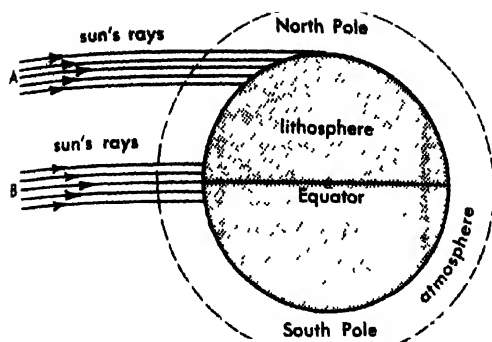


Fig. 2. The oblique ray *A* characteristic of polar regions delivers less energy at the earth's surface than the vertical ray *B* because its energy is spread over a larger surface and because it passes through a thicker layer of absorbing and reflecting atmosphere, which diminishes its energy.

May 11, 1957, a surface temperature of  $-101.5^{\circ}\text{F}$  was observed with clear skies, a 5-mph wind, little compensating heat conduction, absence of sunlight, and an inversion of  $60^{\circ}$  in 2500 ft. However, when a warmer cloud layer moved in, its downward radiation warmed the surface  $16^{\circ}$  in 3 hours.

In the northern polar region, lowest winter temperatures are less coincident with the pole. When appreciable heat conduction from below is available to compensate for the radiative loss, as in the Arctic Basin, surface temperatures do not fall to the low values observed over adjacent continents. The lowest temperatures in the Northern Hemisphere are observed at such continental locations as Verkhoyansk ( $-90^{\circ}\text{F}$ ) in northeastern Siberia. Much farther north, in the Arctic Ocean Basin, the lowest temperature in the pack ice is only  $-62^{\circ}\text{F}$ , and on the thicker Fletcher's Ice Island, a slightly lower  $-65^{\circ}\text{F}$ .

Continental and marine areas differ in temperature regime. The annual march of temperature of a continental polar atmosphere has a sharp minimum, usually one month after the winter solstice. In contrast, the atmosphere over the oceanic Arctic Basin and the Canadian Archipelago has a flat winter minimum of several month's duration during which there may be several minor minima and maxima. Surprisingly enough, these flat winter minima are also found in the decidedly nonmaritime interiors of Greenland and Antarctica. Apparently compensating advection of warmer air from the nearby oceans is responsible and this appears to affect the smaller continental areas, such as Antarctica and Greenland, much more readily than the northern interiors of the larger continents of Asia and North America.

**Water vapor and precipitation.** Low atmospheric temperature decreases water evaporation so that precipitation is meager and fog infrequent. However, where there is open water or artificial release of copious water vapor at a temperature below  $-22^{\circ}\text{F}$ , shallow fogs of minute ice crystals form.

At the South Pole less than 3 in. of water-equivalent accumulates in one year, but the lack of evaporation or runoff enables the build-up of ice, thousands of feet thick over the years.

**Upper air patterns.** Above the shallow surface inversion, air temperature decreases with height until the tropopause is reached. If the lower stratosphere is warm,  $-60$  to  $-75^{\circ}\text{F}$ , then the tropopause inversion is quite sharp. But if the stratosphere is cold,  $-90$  to  $-110^{\circ}\text{F}$ , the tropopause may become ill defined. This disappearance of the winter tropopause was once thought to be characteristic only of the Antarctic atmosphere, but it is now observed when cold, intense cyclones are present in the Arctic stratosphere. High-ranging radiosondes sometimes show what appears to be another tropopause near 18 km. See TROPOPAUSE.

In both polar stratospheres a strong horizontal temperature gradient usually extends across the twilight zone, which in turn causes a strong wind-jet, called the polar-night jet. In the Arctic the degeneration of this jet from a quasi-zonal flow to a meandering, eddying flow often occurs before return of the sun and may result from the large continental-oceanic contrasts in the Northern Hemisphere. This eddying motion, which is associated with marked warming, also redistributes the ozone created in the lower stratosphere by the action of solar ultraviolet radiation on atmospheric oxygen. In contrast, the Antarctic stratospheric jet does not change its zonal character or undergo significant warming until the return of the sun.

**Summer.** The return of the sun's rays to polar regions profoundly changes the temperature, cloud, and wind regimes. The most rapid heating occurs in the lower stratosphere where the solar ultraviolet radiation is strongly absorbed by ozone. In Antarctica the stratospheric heating is so rapid that in two months the winter pattern of strong poleward temperature decrease is reversed. Consequently the intense polar cyclone, with central temperatures near  $-110^{\circ}\text{F}$ , often becomes a weak anticyclone with central temperatures near  $-40^{\circ}\text{F}$ .

Because of the high albedo, or reflectivity, of snow to solar radiation (80–90% of the incident radiation), the amount of absorbed radiation is small, but gradually the returning sun weakens and obliterates the surface inversion. In continental areas where the snow mantle is thin the snow soon melts and exposes the bare ground, which then absorbs practically all of the incident radiation and rapidly warms to temperature values of middle latitudes. In the Arctic, pack ice melts considerably (as much as 3 feet), open-water leads appear, and the albedo decreases to almost 50%. Arctic air temperature remains near  $32^{\circ}\text{F}$ ; but the increased supply of water vapor from evaporation encourages formation of low stratus cloud and fog. In Antarctica, away from exposed mountains, rocks, and other dark surfaces, there is very little snow melt or evaporation. Hence albedo remains high, and weak surface inversions may be found even in summer. Fog and stratus clouds are more

frequent in coastal areas where moist maritime air moves in and is cooled from below.

**Influences of polar air masses.** In winter and summer, polar air masses play an important role in the general circulation of the atmosphere. They move equatorward into and through middle latitudes in sporadic impulses to help equalize the irregular distribution of solar heating over the earth. These outwardly moving polar air masses contribute greatly to the diversity of air, turbulence, and storms which mark the weather and climate of many parts of the middle latitudes. See AIR MASS; FRONT; STORM. [H.W.E.]

**Bibliography:** T. F. Malone (ed.), *Compendium of Meteorology*, 1951; M. P. van Rooy (ed.), *Meteorology of the Antarctic*, 1957.

## Polar molecule

A molecule possessing a permanent electric dipole moment. Molecules containing atoms of more than one element are polar except where forbidden by symmetry; molecules formed from atoms of a single element are nonpolar (except ozone). See DIPOLE MOMENT.

The dipole moments of polar molecules result in stronger intermolecular attraction, increased viscosities, higher melting and boiling points, and greater solubility in polar solvents in comparison with nonpolar molecules.

The electrical response of polar molecules depends in part on their partial alignment in an electric field, the alignment being opposed by thermal agitation forces. This orientation polarization is strongly temperature dependent, in contrast to the induced polarization of nonpolar molecules. See DIELECTRIC CONSTANT; FERROELECTRICS; MOLECULAR STRUCTURE AND SPECTRA; POLARIZATION (DIELECTRICS). [R.D.W.]

## Polar navigation

The complex of navigational techniques modified from those used in other areas to suit the distinctive regional character of polar areas. Although polar navigation has become routine to a rising number of navigators operating in and through such high-latitude parts of the world, their success continues to be based on a sound grasp of the regional differences and the developing adaptations of navigational principles and aids to suit these peculiar area needs. This article therefore singles out salient physical differences, from those of other areas, which demand navigational modifications and relates them to the best known special techniques and applications.

**Coordinates, directions, and charts.** In polar regions, the meridians radiate outward from the poles, and parallels are concentric circles. Thus, the rectangular coordinates familiar to the navigator accustomed to using the Mercator projection are replaced by polar coordinates. A rhumb line is of little use in high latitudes. The usual significance

of direction decreases as latitude increases, disappearing at the poles, where all directions are south (or north).

**Polar navigational grid.** For navigational purposes it is convenient to restore rectangular coordinates. This is done by placing a series of parallel lines on the chart, and letting them represent fictitious meridians. This is called a navigational grid. Generally, these lines are parallel to the Greenwich meridian, and grid north is taken as north along this meridian, as shown in Fig. 1. In this way, grid and true directions near the poles are interconverted by the addition or subtraction of longitude. Variation of the compass (magnetic declination) is replaced by grid variation, or grivation, which is the angle between a magnetic meridian and a grid meridian. Thus, grid and magnetic directions are interconverted by the addition or subtraction of grid variation.

**Selection of chart projection.** As latitude increases, the Mercator chart so familiar to the marine navigator of lower latitudes becomes of decreased utility. In subpolar regions—roughly in the area between the 60 and 70° parallels of latitude—the Lambert conformal projection is used increasingly for nautical charts. In higher latitudes both nautical and aeronautical charts are generally on a polar projection. The polar stereographic and transverse Mercator projections are used most commonly, but the modified Lambert conformal (Ney's), polar gnomonic, and polar azimuthal equidistant projections are also used. See MAP PROJECTIONS.

**Light conditions.** At the poles, the day and year become synonymous as regards the proportion of daylight to darkness. The duration of twilight is lengthened proportionately, lasting several weeks. About 32 hours elapse from the time the lower limb of the sun touches the horizon until the

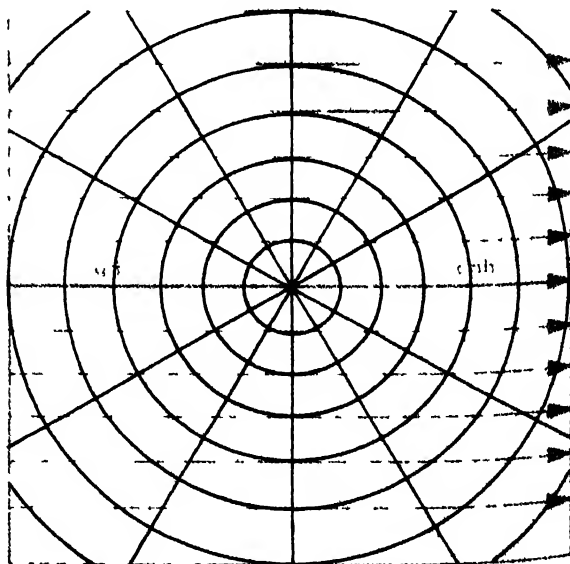


Fig. 1. A polar stereographic grid with a polar grid overprint.

upper limb disappears at sunset, during which time the sun travels  $1\frac{1}{2}$  times ( $480^\circ$ ) around the horizon. The period of several days that occurs each spring and autumn when the sun is below the horizon but too close to it for other celestial bodies to be visible is a critical time for navigation, because of the great dependence upon celestial bodies for both position and direction determination.

**Celestial sphere and time patterns.** The diurnal motion of celestial bodies is essentially horizontal. Half of the celestial sphere is always visible at each pole, and the other half is always below the horizon.

Time loses its usual significance in regions where all time zones come together. The time of any zone would be about equally satisfactory. It is common practice to keep Greenwich mean time.

**Meteorological factors.** Temperatures are generally low, but not as low as might be supposed. It is true that the coldest atmospheric temperatures recorded have been on the Antarctic continent, but temperatures at the North Pole are never as low as some recorded in Yellowstone National Park. Although few places in Antarctica ever get above freezing, much of the Arctic does so regularly each summer. A temperature of  $100^\circ\text{F}$  has been recorded on the Arctic Circle.

The polar air is relatively dry, but because of the low prevailing temperatures, fog and clouds are common. Visibility may also be limited by blowing snow. In most parts of the polar regions, precipitation is so light that these areas are sometimes classed as deserts. Nevertheless, over much of both polar areas the ice does not completely melt during the summer, and large quantities of ice are left both on the land and in the sea. When there is no fog, the atmosphere is exceptionally clear. In the clear, cold air of polar regions sounds travel great distances. It is not unusual to hear the bark of a dog 10 miles away.

The strongest surface winds in the world are probably encountered in certain regions of Antarctica, where speeds of more than 200 knots are not unusual. Over the Arctic Ocean, however, strong winds are not encountered except in some regions near land.

The conditions described are primarily responsible for the differences between navigation in polar regions and elsewhere. There is no well-defined line of demarcation between polar regions and sub-polar regions, nor is there universal agreement as to the definition of polar regions. For the purposes of this discussion, however, the parallel of latitude  $70^\circ$  can conveniently be used as a general dividing line. Polar grid navigation is usually limited to the latitudes poleward of this parallel.

**Piloting hazards.** Piloting in polar regions is strongly affected by the absence of any great number of aids to navigation. Also, natural landmarks may not be shown on the chart, or may be difficult to identify. The appearance of some landmarks changes markedly under different ice conditions. snow covers both the land and a wide ice

foot attached to the shore and extending for miles to seaward, even the shore line is difficult to locate. Adjacent islands sometimes merge together as the straits between them fill up completely with ice. Along a rugged coast such as that of much of Greenland, snow-covered headlands may look alike.

**Unreliability of charts.** Charts of polar regions are less reliable than those of other regions, because relatively little surveying has been done in the polar areas. Attempts to fix the position of a craft can be discouraging when the various landmarks used are not charted in correct position relative to each other. Marine navigators sometimes plot positions relative to land known to be shown in the wrong place, rather than in the correct geographical positions, because it is the land and its attendant shoals that constitute danger to their vessels.

Because relatively few soundings are shown on charts, ships entering harbors often send small boats ahead to determine the depth of water available. The polar marine navigator finds a knowledge of geology useful in predicting safe areas and those in which rocks and shoals might be encountered. Ice concentrations and movement may also be an indication of relative depth of water.

**Limitations of electronic applications.** Electronic aids are not abundant in polar regions. Loran coverage extends to some parts of the Arctic. Radar is useful, but experience in interpretation of the scope in polar regions is essential for reliable results. This is particularly true in aircraft, where the relative appearance of water and land areas often reverse in winter and summer. Hummocked ice presents a different appearance from unbroken ice. A radio direction finder is useful, when radio signals are available. Except along the northern coast of the U.S.S.R., few radio beacons are available. The use of electronics in polar regions is further restricted by magnetic storms, which are particularly severe in the auroral zones.

**Difficulties of dead reckoning.** Reliable dead reckoning depends upon the availability of accurate measurement of direction and distance (or speed). There are difficulties in meeting these requirements in polar regions. See DEAD RECKONING.

**Compass limitations.** Direction is measured largely by a compass. The compasses in common use are the magnetic, in which the directive element attempts to align itself with the horizontal component of the earth's magnetic field, and the north-seeking gyro, in which the directive element attempts to align itself with the earth's axis of rotation. The magnetic compass becomes unreliable in the vicinity of the magnetic poles of the earth, and the north-seeking gyrocompass becomes unreliable in the vicinity of the geographical poles of the earth. The magnetic and geographical poles are both in high latitudes, and the areas in which the compass errors are large and somewhat erratic overlap. In these areas frequent checks and comparisons are needed. The directional gyro is widely used in aircraft operating in polar regions but



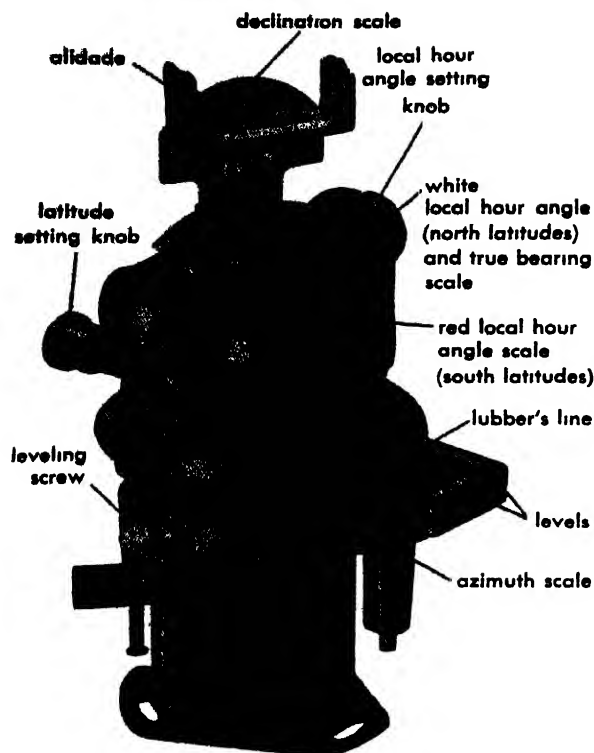


Fig. 2 An astrocompass.

is not in general use aboard ship. See AIRCRAFT COMPASS SYSTEM; GYROSCOPE.

**Celestial determination of direction.** Direction can be determined by means of celestial bodies, but this is usually an instantaneous indication used to check the compasses, unless a device equipped with a clock mechanism is used to provide more continuous information. Several types of devices have been developed to facilitate the use of celestial bodies for determination of direction. The oldest is the sun compass, which utilizes the shadow of a shadow pin, or gnomon, and a suitable dial. This, of course, will not operate unless the sun is visible. A sky compass indicates direction of the sun by means of polarized light in the sky when the sun is near the horizon, even though it may be below the horizon or otherwise obscured. This device may offer the only means of determining direction during the brighter part of the long polar twilight. A Canadian version is called a twilight compass because of the principal period of its use. An astrocompass, shown in Fig. 2, can be set for the coordinates of any celestial body and the latitude of the observer, and then gives an indication of azimuth, true north, and heading.

**Distance or speed problems.** Such measurement in polar regions presents no problems in aircraft. When ships operate in ice, however, the sensing element in the water may be adversely affected or damaged by the ice. A method of determining speed or distance that has proved successful has been to track an iceberg or other prominent feature, either visually or by radar. If the feature is stationary, as

a grounded iceberg, the result is speed or distance over the ground.

At best, dead reckoning is difficult aboard a ship operating in the ice, not only because of the difficulty encountered in measuring course and speed, but also because neither of these may be constant for very long. Land ice which flows down to the sea and breaks off in the form of icebergs or ice islands is not usually a problem, and may even prove beneficial. Its use as an aid in the determination of speed has been mentioned. Individual pieces are usually so large they move with deep-water currents, often in a direction differing from that of the sea ice, which moves mainly in the direction the wind blows. Thus, an iceberg might clear a path in the desired direction of motion. One precaution is essential, however. It is dangerous to approach close to an iceberg, both because of possible under water rams which might extend out for some distance from the berg, and also because it is not uncommon for an iceberg to acquire unstable equilibrium, because of uneven melting, and capsize.

Ice formed at sea, called sea ice, is seldom an unbroken sheet over any very large area. The unequal pressure exerted by tides, currents, winds and temperature changes produce stresses that break the ice and move different parts of it relative to each other, producing leads, long cracks that have opened wide enough to permit passage of a ship; polynyi, large areas, other than leads, relatively free from ice; or pressure ridges, ridges of ice piled up where two floes have come together under pressure. See SEA ICE.

A large field of floating pieces of ice which have drifted together is called the pack. If this is relatively loose, as shown in Fig. 3, a ship, skillfully handled, can negotiate it. If it is packed tightly under pressure, however, it is best avoided.

Successful negotiation of pack ice is the result of working with, not against, the pack, seeking out weak spots, ramming when appropriate, and retreating at other times, taking advantage of leads and polynyi, and avoiding heavy pressure that might

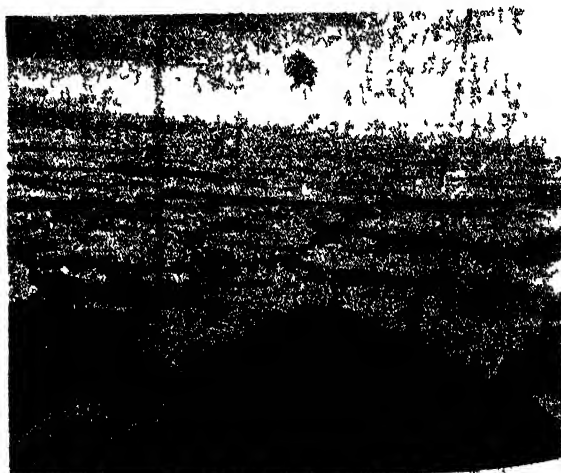


Fig. 3. Pack ice.

stop the vessel or crush it. It is usually considered better to keep moving in the general direction desired than to risk complete stoppage with possible damage by staying exactly on the desired course. All this, plus a lack of detailed information on currents, with little opportunity to acquire it, makes marine dead reckoning difficult in polar regions.

Celestial navigation is of great importance in polar regions, often providing the only means of determining position accurately, or establishing a directional reference.

**Celestial navigation problems.** In aircraft, the use of a sextant is the same as it is in other latitudes. Aboard ship, a marine sextant can be used when celestial bodies and a horizon are both available, an occurrence more frequent in most ocean areas at lower latitudes. Since the pack smooths the sea, resulting in very little rolling or pitching of a ship operating in ice, some marine navigators have found an artificial-horizon sextant of the type carried in aircraft useful. If the acceleration error is too great aboard ship, observations from a nearby ice cake may be possible.

When operating in lower latitudes, navigators generally avoid observation of bodies near the horizon because of the uncertainty of the refraction correction there. In polar regions, even though refraction is more uncertain, navigators often have no choice. Near the equinoxes the sun may be the only body available for several weeks, and it remains close to the horizon. Under these conditions, the navigator observes the sun and makes the most of the information he has, usually with satisfactory results.

During the polar summer, the sun is often the only celestial body available. The moon is above the horizon half the time, but is in a favorable position relative to the sun during relatively few days each month. When only one body is available, it is observed at frequent intervals, perhaps hourly, and a series of running fixes are plotted.

Timing of celestial observations is less critical in polar regions than in lower latitudes because apparent motion of the bodies is nearly horizontal. A relatively large error in longitude, the effect of an error in time, is a relatively small error in miles in an area where the meridians are close together.

Sight reduction in polar regions is not attended by any unusual problems. The same methods in common use in lower latitudes are used in polar regions. Simplified tables might be prepared, but they are generally avoided in favor of familiar methods. See CELESTIAL NAVIGATION; NAVIGATION; PILOTING. [A.B.M.]

**Bibliography:** N. Bowditch, *American Practical Navigator*, U.S. Navy Hydrographic Office, H.O. 9, 1958; U.S. Navy Hydrographic Office, *Air Navigation*, H.O. 216, 1955.

### Polar triangle

The poles of a great circle on a sphere are the points of the sphere where a perpendicular to the of the great circle at its center cuts the

sphere. To obtain the polar triangle of a spherical triangle  $ABC$  (see TRIGONOMETRY, SPHERICAL), construct on its sphere three great circles having respective poles  $A$ ,  $B$ , and  $C$ . Two great circles, one having  $B$  and the other  $C$  as poles, intersect in two points on opposite hemispheres defined by the great circle through  $BC$ . Denote by  $A'$  the point lying on the same hemisphere as  $A$ . Locate  $B'$  and  $C'$  by a like procedure. The spherical triangle  $A'B'C'$  is the polar triangle of  $ABC$ . In geometry it is shown that if one spherical triangle is the polar of a second, then the second is the polar of the first. Also, if the triangles are lettered in the conventional way, then

$$\begin{aligned} A + a' &= B + b' = C + c' = A' + a = B' + b \\ &= C' + c = 180^\circ \end{aligned}$$

[L.M.K.]

### Polar wandering

The large-scale secular movement of the terrestrial poles; the fact of such movement, though based on certain evidence, remains speculative. Such motion is in addition to the small-scale, 14-month cyclic variation (Chandler motion). See GEODESY.

Considerable scientific evidence points toward a relationship between possible polar migration and motion of the earth's subcrustal mantle with respect to the earth's heavy core. If the earth's rigid shell or its various blocks were to "glide" on the moving material beneath, the shell would yield and adjust to changes in the direction of the polar axis.

Scientific discussion has been animated over the theory of westward drift of the New World continents in relation to the Old World continents as introduced by A. Wegener in the 1920s (as also in ocean basin formation, see SUBMARINE TOPOGRAPHY). Many scientists now feel that the rate of such drift must be many times smaller than assumed by Wegener. Some believe that it is almost a proved fact that the Atlantic Ocean basin will become broader—whether from westward drift of the New World continents, from the Old World continents moving eastward, or from both moving eastward but at a higher rate eastward for the Old World continents. To many the last seems most likely and fits theories of adjustment to internal convection movements and to shifts in polar axis.

Other conditions contribute more or less directly to the concept of polar wandering. Vertical and horizontal shifts in continental and oceanic blocks may be in some circumstances not only an adjustment to shifts in polar axis but also contribute to further shifts in the axis. The polar-flight theories of centrifugal force tending to bring the continental shields toward the Equator appear to lack sufficient force to cause axial shifting. Theories of internal movement, however, do allow reasonable explanations for growing divergence, from an earlier coincidence, in position between magnetic and geographic poles. Relative deformation in the material between the crust and core can also explain sev-

eral characteristics of the earth's magnetic field, such as the north drift and counterclockwise rotation of India relative to the poles. Studies in geology and rock magnetism give evidence of past differences in location of the magnetic poles and of possibilities for location of the poles—for example the North Pole in Alaska during the Tertiary, and during earlier geological times in India, as Vening Meinesz assumes (see ROCK MAGNETISM). However, observation and analysis of present data are insufficient to warrant considering that the hypothesis of polar wandering is completely established. For a discussion of nutational wandering of the terrestrial pole, see NUTATION (ASTRONOMY AND MECHANICS). [W.A.H.]

## Polar-coordinate navigation systems

Systems in which one or more signals are emitted from a facility (or co-located facilities) to produce simultaneous indication of bearing and distance, also called Rho-theta systems. Since a bearing is a radial line of position and a distance is a circular line of position, the polar-coordinate system always insures a position fix produced by the intersection of two lines of position which are at right angles to each other. Since the reference for both lines of position is at a common origin, computation for any course referred to this origin is simplified. See NAVIGATION SYSTEMS, ELECTRONIC. [P.C.S.]

## Polarimetric analysis

A method of chemical analysis based on the optical activity of the substance being determined. Optically active materials are asymmetric, that is, their molecules or crystals have no plane or center of symmetry. These asymmetric molecules can occur in either of two forms, *d*- and *l*-, called optical isomers. Often a third optically inactive form, called meso, also exists. Asymmetric substances possess the power of rotating the plane of polarization of plane-polarized light. Measurement of the extent of this rotation is called polarimetry. Polarimetry is applied to both organic and inorganic materials. See OPTICAL ACTIVITY.

The extent of the rotation depends on the character of the substance, the length of the light path, the temperature of the solution, the wavelength of the light which is being used, the solvent (if there is one), and the concentration of the substance. In most work, the yellow light of the D line of the so-

dium spectrum (5893 Å) is used to determine the specific rotation.

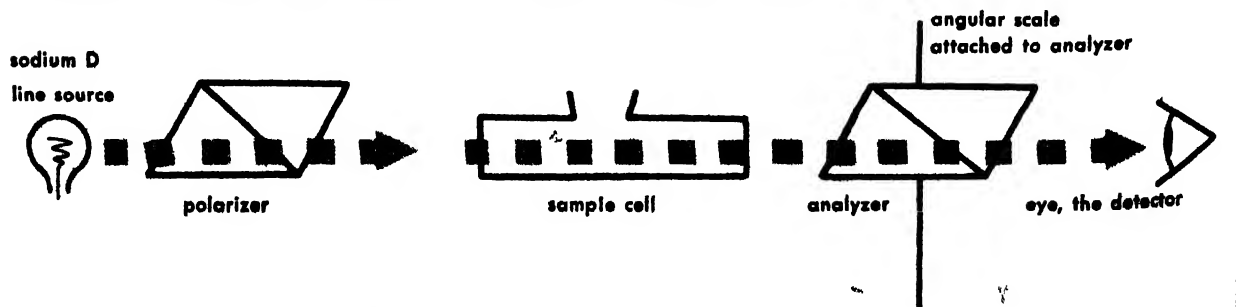
$$\text{Specific rotation} = [\alpha]_D^{20} = \frac{\alpha}{l\rho}$$

where  $\alpha$  is the measured angle of rotation,  $l$  the length of the column of liquid in decimeters, and  $\rho$  the density of the solution. In other words, the specific rotation is the rotation in degrees which this plane-polarized light of the sodium D line undergoes in passing through a 10 cm long sample tube containing a solution of 1 g/ml concentration at 20°C.

In the simplified diagram of a polarimeter light from the sodium lamp is polarized by the polarizer (a fixed nicol prism) before it passes through the cell containing the material being analyzed. After the light passes through the cell, it passes through the analyzer (another nicol prism) and then is detected by the eye or a photocell. A comparison of the angular orientation of the analyzer as measured on the scale with the cell empty and with the cell filled with solution serves to measure the rotation of the polarized light by the sample. This rotation may be either clockwise (+) or counterclockwise (−), depending on the substance in question.

Polarimetry may be used for either qualitative or quantitative analytical work. In qualitative applications, the presence of an optically active material is shown, and then a calculation of specific rotation often leads to the identification of the unknown. In quantitative work, the concentration of a given optically active material is determined either from a calibration curve of percentage of the constituent versus angular rotation or from the specific rotation, assuming the angular rotation to be a linear function of concentration. For this method of analysis to be useful, it is necessary that only one optically active material be present in solution.

Polarimetry is widely used in carbohydrate chemistry, especially in the analysis of sugar solutions. Polarimeters used for this work are specially designed and are called saccharimeters. Other materials often determined by polarimetry are tartaric acid, Rochelle salt (potassium sodium tartrate), various terpenes such as *d*- and *l*-pinene, many steroids, and other compounds of biological and biochemical importance. Since there is great difference between the biological activities of the different optical forms of organic compounds, polarimetry is widely



Simplified diagram of a polarimeter.

used in biochemical research to identify the molecular configurations.

Optical rotatory dispersion is the measurement of the specific rotation as a function of wavelength. To accomplish this, the sodium lamp is replaced by a monochromator, and a source of continuous radiation. A photocell circuit is substituted for the eye as a detector. In this way, the specific rotation may be determined in the ultraviolet or near infrared, as well as in the visible region of the spectrum.

The information obtained by optical rotatory dispersion has shown that minor changes in configuration of a molecule have a marked effect on its dispersion properties. By using the properties of compounds of known configuration, it has been possible to determine the absolute configurations of many other molecules and to identify various isomers. To date, most of the applications have been to steroids, sugars, and other natural products. See OPTICAL METHODS OF CHEMICAL ANALYSIS; POLARIZED LIGHT; ROTATORY DISPERSION. [R.F.G.]

## Polaris

Alpha Ursae Minoris, the North Star, or pole star. Polaris is located approximately  $1^\circ$  of arc from the north celestial pole, the point where the Earth's axis of rotation now pierces the celestial sphere. Two bright stars at the end of the bowl of the Big Dipper point toward Polaris.

Polaris is a high-luminosity star of spectral type late F. It is one of the brightest cepheid variables, but is atypical in having the small light range of only 0.1 magnitude. The period is 4 days, the luminosity 3000 times that of the Sun, distance 200 parsecs. Polaris has a faint, main-sequence companion. See STAR. [J.L.GR.]

## Polarity (electricity)

A term which refers to the designation of positive and negative terminals in an electric circuit carrying a current. The choice of positive charge, and thereby of positive potential, is a purely arbitrary one which had been made long before the discovery of the electron. It is now known that the primary current carriers in most circuit elements are electrons, which possess negative charge.

For a seat of electromotive force, such as a battery or a generator, the positive terminal is by convention the one at which electrons enter from the external circuit; the negative terminal is the one from which they leave. For a "passive" element, such as an ammeter or a resistor, the positive terminal is the one through which the electrons leave the element; the negative terminal is the one through which they enter from the circuit. The potential of the positive terminal of any element is by definition greater in value than that of the negative terminal. In all cases the conventional current flows in the opposite direction to the electrons.

In the case of alternating current the term polarity may be used in a different sense. Usually in commercial circuits one leg is maintained at zero po-

tential while the other is alternately above and below ground. In some electrical appliances, it makes a difference which side is connected to the "hot" lead. Turning the plug around in the wall socket to obtain the best operation of a particular device is often referred to as "reversing the polarity."

[J.W.ST.]

## Polarization (dielectrics)

A vector quantity representing the electric dipole moment per unit volume of a dielectric material. See DIELECTRICS; DIPOLE MOMENT.

The polarization  $P$  is related to the macroscopic electric parameters by the equations

$$D = \epsilon_0 E + \epsilon_0(\kappa' - 1)E = \chi \epsilon_0 E$$

where  $D$  is the electric displacement,  $E$  is the electric field strength,  $\epsilon_0$  is the permittivity of vacuum,  $\kappa'$  is the dielectric constant,  $\chi$  is the electric susceptibility, and  $\gamma$  is a geometrical factor. In cgs electrostatic units  $\epsilon_0 = 1$  and  $\gamma = 4\pi$ ;  $\epsilon_0 = 8.854 \times 10^{-12}$  farad/m and  $\gamma = 1$  in rationalized mks units. The dimensions of polarization are statcoulomb per square centimeter in the cgs system and coulomb per square meter in the rationalized mks system. See ELECTRICAL UNITS.

Dielectric polarization arises from the electrical response of individual molecules of a medium and may be classified as electronic, atomic, orientation, and space-charge or interfacial polarization, according to the mechanism involved.

Electronic polarization represents the distortion of the electron distribution or motion about the nuclei in an electric field. This polarization occurs for all materials and is nearly independent of temperature and frequency up to about  $10^{14}$  cps for insulators.

Atomic polarization arises from the change in dipole moment accompanying the stretching of chemical bonds between unlike atoms in molecules. This mechanism contributes to polarization at frequencies below those of the vibrational modes of molecules (about  $10^{12}$ – $10^{14}$  cps). For a discussion of molecular vibrations, see MOLECULAR STRUCTURE AND SPECTRA.

Orientation polarization is caused by the partial alignment of polar molecules, that is, molecules possessing permanent dipole moments, in an electric field. This mechanism leads to a temperature-dependent component of polarization at lower frequencies.

Space-charge or interfacial polarization occurs when charge carriers are present which can migrate an appreciable distance through a dielectric but which become trapped or cannot discharge at an electrode. This process always results in a distortion of the macroscopic field and is important only at low frequencies.

See DIELECTRIC CONSTANT; ELECTRIC FIELD; SUSCEPTIBILITY, ELECTRIC. [R.D.W.]

## Polarization of waves

Polarization is the phenomenon which is exhibited when a transverse wave is polarized. The term polarization is also used to describe the process of polarizing a wave.

In an unpolarized wave, the vibrations in a plane perpendicular to the ray appear to be oriented in all directions with equal probability. In a polarized wave the displacement direction of the vibrations is completely predictable. For certain disturbances, such as the transverse acoustic wave produced when a steel bar is struck, the polarization is complete. Electromagnetic radiation is normally unpolarized if it is generated by atomic processes (see ELECTROMAGNETIC RADIATION). Thus ultraviolet, visible, and infrared radiations produced by heated bodies or electrical discharges are generally unpolarized. Radiation generated by vacuum-tube oscillators or transistor oscillators is always polarized. The probability waves (matter waves) associated with atomic or nuclear particles are generally unpolarized. See QUANTUM MECHANICS.

Some of the different types of polarization, as well as the technique of producing polarization in an unpolarized wave, are described in another article (see POLARIZED LIGHT). The electric vector can lie in a plane or it can follow a path whose projection at right angles to the direction of propagation is a circle or an ellipse. The same types of polarization can be produced in any transverse wave. For example, see MICROWAVE OPTICS.

Electromagnetic radiation is difficult to polarize in certain spectral regions, and few techniques exist for analysis. This is true in the ultraviolet below 1900 Å. No dichroic polarizers have been found for this region, and transparent birefringent materials from which Nicol or Wollaston polarizing prisms could be made do not seem to exist. Polarization by reflection is possible, but very little work has been done with this technique. In the infrared region from the end of the visible spectrum to approximately  $2\mu$ , sheet polarizers exist. To around  $4\mu$ , polarizing prisms can be made. From  $4\mu$  to  $80\mu$ , reflection from a single plate or transmission through a pile of transparent plates is the common procedure. All these techniques produce linear polarization. Elliptical or circular polarization is more difficult to achieve.

X-ray photons, electrons, neutrons, and other particles can be polarized most easily by scattering. [B.H.BI.]

## Polarized light

Light which has its electric vector oriented in a predictable fashion with respect to the propagation direction. In unpolarized light, the vector is oriented in a random, unpredictable fashion. Even in short time intervals, it appears to be oriented in all directions with equal probability. Most light sources seem to be partially polarized so that some fraction of the light is polarized and the remainder

unpolarized. It is actually more difficult to produce a completely unpolarized beam of light than one which is completely polarized.

The polarization of light differs from its other properties in that human sense organs are essentially unable to detect the presence of polarization. The Polaroid Corporation with its polarizing sunglasses and camera filters has made millions of people conscious of phenomena associated with polarization. Light from a rainbow is completely linearly polarized, that is, the electric vector lies in a plane. The possessor of polarizing sunglasses discovers that with such glasses, the light from a section of the rainbow is extinguished.

According to all available theoretical and experimental evidence, it is the electric vector rather than the magnetic vector of a light wave that is responsible for all the effects of polarization and other observed phenomena associated with light. Therefore, the electric vector of a light wave, for all practical purposes, can be identified as the light vector. See ELECTROMAGNETIC RADIATION; LIGHT, POLARIZATION OF WAVES. For information which is closely related to much of the ensuing discussion see CRYSTAL OPTICS.

One of the simplest ways of producing linearly polarized light is by reflection from a dielectric surface. At a particular angle of incidence, the reflectivity for light whose electric vector is in the plane of incidence becomes zero. The reflected light is thus linearly polarized at right angles to the plane of incidence. This fact was discovered by E. Malus in 1808. Brewster's law shows that at the polarizing angle the refracted ray makes an angle of  $90^\circ$  with the reflected ray. By combining this relation ship with Snell's law of refraction, one finds that

$$\tan i = n \quad (1)$$

where  $i$  is the angle of incidence and  $n$  is the refractive index. This provides a simple way of measuring refractive indices. See REFRACTION OF WAVES.

**Law of Malus.** If linearly polarized light is incident on a dielectric surface at Brewster's angle (the polarizing angle), then the reflectivity of the surface will depend on the angle between the incident electric vector and the plane of incidence. When the vector is in the plane of incidence, the reflectivity will be zero. When it is at right angles, the reflectivity will be at a maximum. To compute the complete relationship, the incident light vector  $A$  is broken into components, one vibrating in the plane of incidence and one at right angles to the plane:

$$A_{\parallel} = A \sin \theta \quad (2)$$

$$A_{\perp} = A \cos \theta \quad (3)$$

where  $\theta$  is the angle between the light vector and a plane perpendicular to the plane of incidence. Since the component in the plane of incidence is not reflected, the reflected ray can be written

$$B = A \cos \theta \quad (4)$$

where  $r$  is the reflectivity at Brewster's angle. The intensity is

$$I = B^2 = A^2 r^2 \cos^2 \theta \quad (5)$$

This is the mathematical statement of the law of Malus.

**Linear polarizing devices.** The angle  $\theta$  can be considered as the angle between the transmitting axes of a pair of linear polarizers. When the polarizers are parallel, they are transparent. When they are crossed, the combination is opaque. The first polarizers were glass plates inclined so that the incident light was at Brewster's angle. Such polarizers are quite inefficient since only a small percentage of the incident light is reflected as polarized light. More efficient polarizers can be constructed.

**Dichroic crystals.** Certain natural materials absorb linearly polarized light of one vibration direction much more strongly than light vibrating at right angles. Such materials are termed dichroic and are described more fully in another article (see DICHROISM). Tourmaline is one of the best known dichroic crystals, and tourmaline plates were used as polarizers for many years. A pair was usually mounted in what were known as tourmaline tongs.

**Birefringent crystals** Other natural materials exist in which the velocity of light depends on the vibration direction. These materials are called birefringent. The simplest of these structures are crystals in which there is one direction of propagation for which the light velocity is independent of its state of polarization. These are termed uniaxial crystals, and the propagation direction mentioned is called the optic axis. For all other propagation directions, the electric vector can be split into two components, one lying in a plane containing the optic axis and the other at right angles. The light velocity or refractive index for these two waves is different.

One of the best known of these birefringent crystals is transparent calcite (Iceland spar), and a series of polarizers have been made from this substance. W. Nicol (1829) invented the Nicol prism, which is made of two pieces of calcite cemented together as in Fig. 1. The cement is Canada balsam, in which the wave velocity is intermediate between the velocity in calcite for the fast and the slow ray. The angle at which the light strikes the boundary is such that for one ray the angle of incidence



Fig. 1. Nicol prism. The ray for which Snell's law holds is called the ordinary ray.

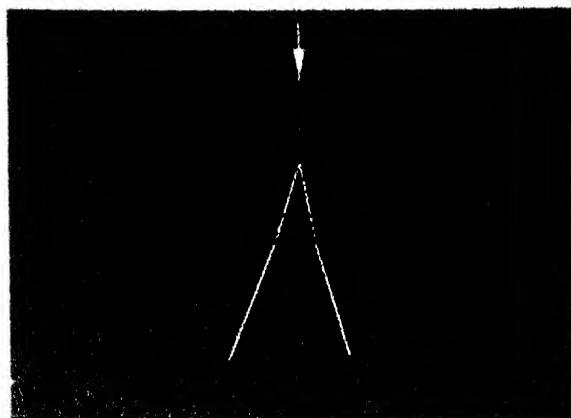


Fig. 2. Wollaston prism.

is greater than the critical angle for total reflection. Thus the rhomb is transparent for only one polarization direction.

Canada balsam is not completely transparent in the ultraviolet at wavelengths shorter than 4000 angstroms. Furthermore, large pieces of calcite material are exceedingly rare. A series of polarizers has been made using quartz, which is transparent in the ultraviolet and which is more commonly available in large pieces. Because of the small difference between the refractive indices of quartz and Canada balsam, a Nicol prism of quartz would be tremendously long for a given linear aperture.

A different type of polarizer, made of quartz, was invented by W. H. Wollaston and is shown in Fig. 2. Here the vibration directions are different in the two pieces so that the two rays are deviated as they pass through the material. The incoming light beam is thus separated into two oppositely linearly polarized beams which have an angular separation between them. By using appropriate optical stops (obstacles that restrict light rays) in the system, one can select either beam.

In the Wollaston prism, both beams are deviated; and since the quartz produces dispersion, each beam is spread into a spectrum. This is not the case in a prism which was invented by A. Rochon. Here the two pieces are arranged as in Fig. 3. One beam proceeds undeviated through the device and is thus achromatic.

**Sheet polarizers.** A third mechanism for obtaining polarized light is the Polaroid sheet polarizer invented by E. H. Land. Sheet polarizers fall into three types. The first is a microcrystalline polarizer in which small crystals of a dichroic material are oriented parallel to each other in a plastic medium. Typical microcrystals, such as needle-shaped quinine iodosulfate, are embedded in a viscous plastic and are oriented by extruding the material through a slit.

The second type depends for its dichroism on a property of an iodine-in-water solution. The iodine appears to form a linear high polymer. If the iodine is put on a transparent oriented sheet of material such as polyvinyl alcohol (PVA), the iodine



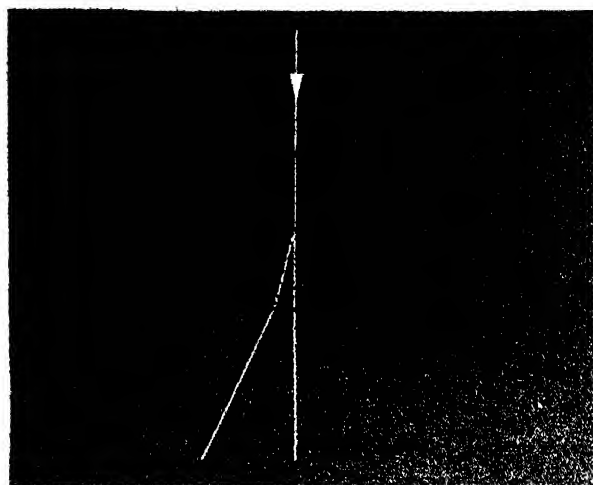


Fig. 3. Rochon prism.

chains apparently line themselves parallel to the PVA molecules and the resulting dyed sheet is strongly dichroic. A third type of sheet polarizer depends for its dichroism directly on the molecules of the plastic itself. This plastic consists of oriented polyvinylene. Because these polarizers are commercially available and can be obtained in large sheets, many experiments involving polarized light have been performed which would have been quite difficult with the reflection polarizers or the birefringent crystal polarizer. See VECTOGRAPH.

**Characteristics.** There are several characteristics of linear polarizers which are of interest to the experimenter. First is perhaps the transmission for light polarized parallel and perpendicular to the axis of the polarizer; second is the angular field; and third, the linear aperture. A typical sheet polarizer has a transmittance of 48% for light parallel to the axis and  $2 \times 10^{-4}\%$  for light perpendicular to the axis at a wavelength of  $550 \text{ m}\mu$ . The angular field is  $60^\circ$ , and sheets can be many feet in diameter. The transmittance perpendicular to the axis varies over the angular field.

The Nicol prism has transmittance similar to that of the Polaroid sheeting, but a much reduced linear and angular aperture.

**Polarization by scattering.** When an unpolarized light beam is scattered by molecules or small particles, the light observed at right angles to the original beam is polarized. The light vector in the original beam can be considered as driving the equivalent oscillators (nuclei and electrons) in the molecules. There is no longitudinal component in the original light beam. Accordingly, the scattered light observed at right angles to the beam can only be polarized with the electric vector at right angles to the propagation direction of the original beam. In most situations, the scattered light is only partially polarized because of multiple scattering. The best known example of polarization by scattering is the light of the north sky. The percentage polarization can be quite high in clean country air. The late A. H. Pfund invented a technique for using

measurements of sky polarization to determine the position of the sun when it is below the horizon. See SCATTERING (ELECTROMAGNETIC RADIATION).

**Types of polarized light.** Polarized light is classified according to the orientation of the electric vector. In linearly polarized light, the electric vector remains in a plane containing the propagation direction. For monochromatic light, the amplitude of the vector changes sinusoidally with time. In circularly polarized light, the tip of the electric vector describes a circular helix about the propagation direction. The amplitude of the vector is constant. The frequency of rotation is equal to the frequency of the light. In elliptically polarized light, the vector also rotates about the propagation direction, but the amplitude of the vector changes so that the projection of the vector on a plane at right angles to the propagation direction describes an ellipse.

These different types of polarized light can all be broken down into two linear components at right angles to each other.

$$E_x = A_x \sin(\omega t + \varphi_x) \quad (6)$$

$$E_y = A_y \sin(\omega t + \varphi_y) \quad (7)$$

where  $A_x$  and  $A_y$  are the amplitudes,  $\varphi_x$  and  $\varphi_y$  the phases,  $\omega$  is  $2\pi$  times the frequency, and  $t$  is the time. For linearly polarized light

$$\varphi_x = \varphi_y \quad A_x \neq A_y$$

For circularly polarized light

$$\varphi_x = \varphi_y \pm \frac{\pi}{2} \quad A_x = A_y$$

For elliptically polarized light

$$\varphi_x \neq \varphi_y \quad A_x \neq A_y$$

In the last case, it is always possible to find a set of orthogonal axes inclined at an angle  $\alpha$  to  $x$  and  $y$  along which the components will be  $E'_x$  and  $E'_y$ , such that

$$\varphi'_x = \varphi'_y \pm \frac{\pi}{2} \quad \text{and} \quad A'_x \neq A'_y$$

In this new system, the  $x'$  and  $y'$  amplitudes will be the major and minor axes  $a$  and  $b$  of the ellipse described by the light vector and  $\alpha$  will be the angle of orientation of the ellipse axes with respect to the original coordinate system. The relationships between the different quantities can be written

$$\tan 2\alpha = \tan 2\gamma \cos \varphi \quad (8)$$

$$\sin 2\beta = \sin 2\gamma \sin \varphi \quad (9)$$

where

$$\tan \gamma = A_y / A_x \quad (10)$$

$$\varphi = \varphi_x - \varphi_y \quad (11)$$

$$\tan \beta = \pm b/a \quad (12)$$

$$A_x^2 + A_y^2 = a^2 + b^2 \quad (13)$$

These same types of polarized light can also be broken down into right and left circular components or into two orthogonal elliptical components

These different vector bases are useful in different physical situations.

**Production of polarized light.** Linear polarizers have already been discussed. Circularly and elliptically polarized light are normally produced by combining a linear polarizer with a wave plate. A Fresnel rhomb can be used to produce circularly polarized light.

**Wave plate.** A plate of material which is linearly birefringent is called a wave plate or retardation sheet. Wave plates have a pair of orthogonal axes which are designated fast and slow. Polarized light with its electric vector parallel to the fast axis travels faster than light polarized parallel to the slow axis. The thickness of the material can be chosen so that for light traversing the plate, there is a definite phase shift between the fast component and the slow component. A plate with a  $90^\circ$  phase shift is termed a quarter-wave plate. The retardation in waves is given by the expression

$$\delta = \frac{(n_s - n_f)d}{\lambda} \quad (14)$$

where  $n_s - n_f$  is the birefringence;  $n_s$  is the slow index at wavelength  $\lambda$ ;  $n_f$  is the fast index; and  $d$  is the plate thickness.

Wave plates can be made by preparing X-cut sections of quartz, calcite, or other birefringent crys-

tals. For retardations of less than a few waves, it is easiest to use sheets of oriented plastics or of split mica. A quarter-wave plate for the visible or infrared is easy to fabricate from mica. The plastic wrappers from many American cigarette packages seem to have almost exactly a half-wave retardation for green light. Since mica is not transparent in the ultraviolet, a small retardation in this region is most easily achieved by crossing two quartz plates which differ by the requisite thickness.

Linearly polarized light incident normally on a quarter-wave plate and oriented at  $45^\circ$  to the fast axis can be split into two equal components parallel to the fast and slow axes. These can be represented, before passing through the plate, by the equations

$$E_x = A_x \sin(\omega t + \varphi_x) \quad (15)$$

$$E_y = A_x \sin(\omega t + \varphi_x) \quad (16)$$

where  $x$  and  $y$  are parallel to the wave-plate axes. After passing through the plate, the two components can be written

$$E_x = A_x \sin\left(\omega t + \varphi_x + \frac{\pi}{2}\right) \quad (17)$$

$$E_y = A_x \sin(\omega t + \varphi_x) \quad (18)$$

where  $E_x$  is now advanced one quarter-wave with respect to  $E_y$ .

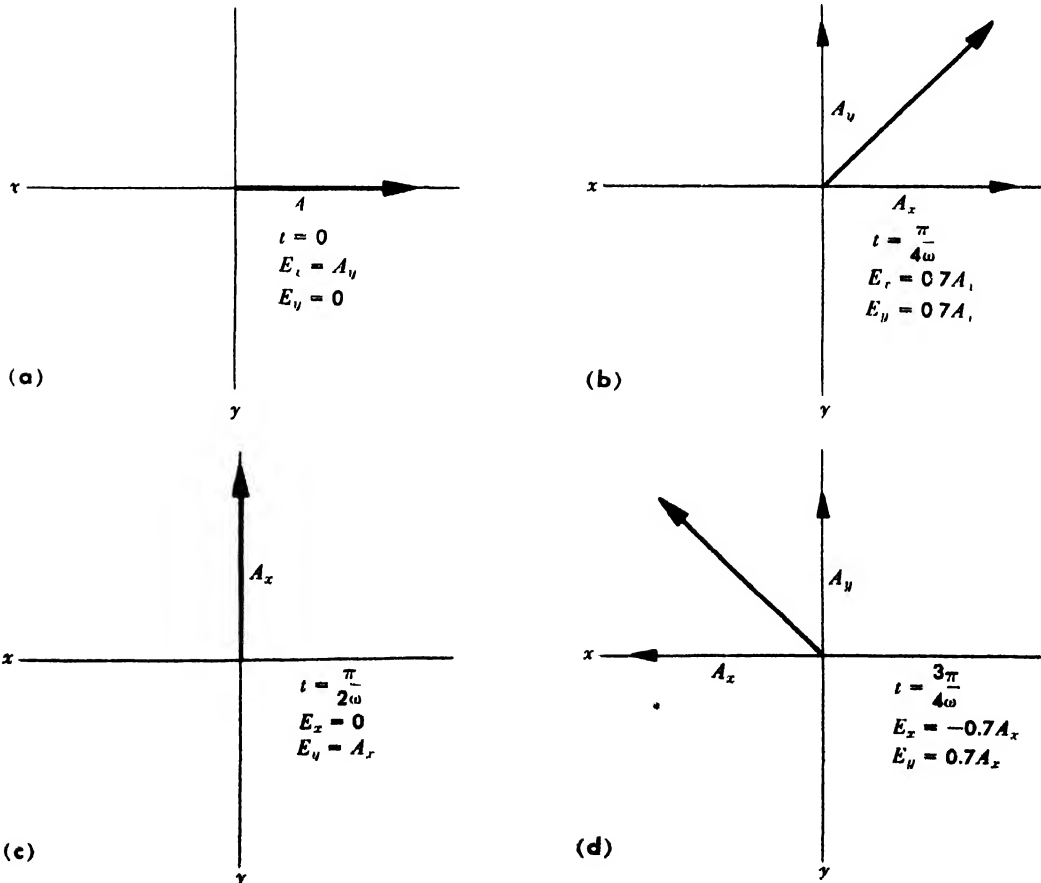


Fig. 4. Projection of light vector on plane  $z = 0$  for circularly polarized light. (a)  $t = 0$ . (b)  $t = \pi/4\omega$ . (c)

$t = \pi/2\omega$ . (d)  $t = 3\pi/4\omega$ . The light vector is of constant amplitude.

One can visualize the behavior of the light by studying the sketches in Fig. 4, which show the projection on a plane  $z = 0$  at various times. It is apparent that the light vector is of constant amplitude, and that the projection on a plane normal to the propagation direction is a circle. If the linearly polarized light is oriented at  $-45^\circ$  to the fast axis, the light vector will revolve in the opposite direction. Thus it is possible with a quarter-wave plate and a linear polarizer to make either right or left circularly polarized light. If the linearly polarized light is at an angle other than  $45^\circ$  to the fast axis, the transmitted radiation will be elliptically polarized. When circularly polarized light is incident on a quarter-wave plate, the transmitted light is linearly polarized at an angle of  $45^\circ$  to the wave-plate axes. This polarization is independent of the orientation of the wave-plate axis. For elliptically polarized light, the behavior of the quarter-wave plate is much more complicated. However, as was mentioned earlier, the elliptically polarized light can be considered as composed of two linear components parallel to the major and minor axes of the ellipse and with a quarter-wave phase difference between them. If the quarter-wave plate is oriented parallel to the axes of the ellipse, the two transmitted components will either have zero phase difference or a  $180^\circ$  phase difference and will be linearly polarized. At other angles, the transmitted light will still be elliptically polarized, but with different major and minor axes. Similar treatment for a half-wave plate shows that linearly polarized light oriented at an angle  $\theta$  to the fast axis is transmitted as linearly polarized light oriented at an angle  $-\theta$  to the fast axis.

Wave plates all possess a different retardation at each wavelength. This appears immediately from Eq. (14). It is conceivable that a substance could have dispersion of birefringence, such as to make the retardation of a plate independent of wavelength. However, no material having such a characteristic has as yet been found.

**Fresnel rhomb.** A quarter-wave retardation can be provided achromatically by the Fresnel rhomb. This device depends on the phase shift which occurs at total internal reflection. When linearly polarized light is totally internally reflected, it experiences a phase shift which depends on the angle of reflection, the refractive index of the material, and the orientation of the plane of polarization. Light polarized in the plane of incidence experiences a phase shift which is different from that of light polarized at right angles to the plane of incidence. Light polarized at an intermediate angle can be split into two components, parallel and at right angles to the plane of incidence, and the two components mathematically combined after reflection.

The phase shifts can be written

$$\tan \frac{\varphi_{\parallel}}{2} = \frac{n \sqrt{n^2 \sin^2 i - 1}}{\cos i} \quad (19)$$

$$\tan \frac{\varphi_{\perp}}{2} = \frac{\sqrt{n^2 \sin^2 i - 1}}{n \cos i} \quad (20)$$

where  $\varphi_{\parallel}$  is the phase shift parallel to the plane of incidence;  $\varphi_{\perp}$  is the phase shift at right angles to the plane of incidence;  $i$  is the angle of incidence on the totally reflecting internal surface; and  $n$  is the refractive index. The difference  $\varphi_{\parallel} - \varphi_{\perp}$  reaches a value of about  $\pi/4$  at an angle of  $52^\circ$  for  $n = 1.50$ . Two such reflections give a retardation of  $\pi/2$ . The Fresnel rhomb shown in Fig. 5 is cut so that the incident light is reflected twice at  $52^\circ$ . Accordingly, light polarized at  $45^\circ$  to the principal plane will be split into two equal components which will be shifted a quarter-wave with respect to each other, and the transmitted light will be circularly polarized. Nearly achromatic wave plates can be made by using a series of wave plates in series with their axes oriented at different specific angles with respect to a coordinate system.

**Analyzing devices.** Polarized light is one of the most useful tools for studying the characteristics of materials. The absorption constant and refractive index of a metal can be calculated by measuring the effect of the metal on polarized light reflected from its surface. See REFLECTION (FETTER, TROMAGNETIC RADIATION).

The analysis of polarized light can be performed with a variety of different devices. If the light is linearly polarized, it can be extinguished by a linear polarizer and the direction of polarization of the light determined directly from the orientation of the polarizer. If the light is elliptically polarized, it can be analyzed with the combination of a quarter-wave plate and a linear polarizer. Any such combination of polarizer and analyzer is called a polariscope. As explained previously, a quarter-wave plate oriented parallel to one of the axes of the ellipse will transform elliptically polarized light to linearly polarized light. Accordingly, the quarter-wave plate is rotated until the beam is extinguished by the linear polarizer. At this point the orientation of the quarter-wave plate gives the orientation of the ellipse and the orientation of the polarizer gives the ratio of the major to the minor axis. Knowledge of the origin of the elliptically polarized light usually gives the orientation of the components which produced it, and from these various items, the phase shifts and attenuations produced by the experiment can be deduced.

One of the best-known tools for working with polarized light is the Babinet compensator. This device is normally made of two quartz prisms put together in a rhomb. One prism is cut with the optic axis in the plane of incidence of the prism, and the

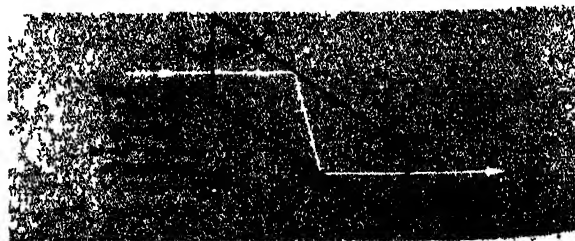


Fig. 5. Fresnel rhomb.

other with the optic axis perpendicular to the plane of incidence. The retardation is a function of distance along the rhomb; it will be zero at the center, varying to positive and negative values in opposite directions along the rhomb. It can be used to cancel the known or unknown retardation of any wave plate.

**Retardation theory.** It is difficult to see intuitively the effect of a series of retardation plates or even of a single plate of general retardation  $\delta$  on light which is normally incident on the plate and which is polarized in a general fashion. This problem is most easily solved algebraically. The single-wave plate is assumed to be oriented normal to the direction of propagation of the light, which is taken to be the  $z$  direction of a set of Cartesian coordinates. Its fast axis is at an angle  $\alpha$  to the  $x$  axis. The incident light can be represented by the equations

$$E_x = A_x \sin(\omega t + \varphi_x) \quad (21)$$

$$E_y = A_y \sin(\omega t + \varphi_y) \quad (22)$$

A first step is to break the light up into components  $E'_x$  and  $E'_y$  parallel to the axes of the plate. One can write

$$E'_x = E_x \cos \alpha - E_y \sin \alpha \quad (23)$$

$$E'_y = E_x \sin \alpha + E_y \cos \alpha \quad (24)$$

These components can also be written

$$E'_x = A'_x \sin(\omega t + \varphi'_x) \quad (25)$$

$$E'_y = A'_y \sin(\omega t + \varphi'_y) \quad (26)$$

After passing through the plate, the components become

$$E''_x = A'_x \sin(\omega t + \varphi'_x + \delta) \quad (27)$$

$$E''_y = A'_y \sin(\omega t + \varphi'_y) \quad (28)$$

In general, it is of interest to compare the output with the input. The transmitted light is thus broken down into components along the original axes. This results in the equations

$$E'''_x = E''_x \cos \alpha + E''_y \sin \alpha \quad (29)$$

$$E'''_y = -E''_x \sin \alpha + E''_y \cos \alpha \quad (30)$$

With this set of equations, it is possible to compute the effect of a wave plate on any form of polarized light.

**Jones calculus.** Equations (29) and (30) still become overwhelmingly complicated in any system involving several optical elements. Various methods have been developed to simplify the problem and enable one to make some generalizations about systems of elements. One of the most straightforward, proposed by R. C. Jones, involves reducing Eqs (29) and (30) to matrix form. The Jones calculus for optical systems involves the polarized electric components of the light vector and is distinguished from other methods in that it takes cognizance of the absolute phase of the light wave.

The Jones calculus writes the light vector in complex form

$$\begin{aligned} E_x &= A_x e^{i(\omega t + \varphi_x)} \\ E_y &= A_y e^{i(\omega t + \varphi_y)} \end{aligned} \quad (31)$$

or

$$E = \begin{pmatrix} A_x e^{i\varphi_x} \\ A_y e^{i\varphi_y} \end{pmatrix} e^{i\omega t} \quad (32)$$

Matrix operators are developed for different optical elements. From Eqs. (21) to (30), the operator for a wave plate can be derived directly. See MATRIX THEORY.

The Jones calculus is ordinarily used in a normalized form which simplifies the matrices to a considerable extent. In this form, the terms involving the actual amplitude and absolute phase of the vectors and operators are factored out of the expressions. The intensity of the light beam is reduced to unity in the normalized vector so that

$$A_x^2 + A_y^2 = 1 \quad (33)$$

Under this arrangement, the matrices for various types of operations can be written

$$G(\delta) = \begin{vmatrix} e^{i(\delta/2)} & 0 \\ 0 & e^{-i(\delta/2)} \end{vmatrix} \quad (34)$$

This is the operator for a wave plate of retardation  $\delta$  and with axes along  $x$  and  $y$ .

$$S(\alpha) = \begin{vmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{vmatrix} \quad (35)$$

This is the operator for a rotator which rotates linearly polarized light through an angle  $\alpha$ .

$$P_h = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} \quad (36)$$

This is the operator for a perfect linear polarizer parallel to the  $x$  axis. A wave plate at an angle  $\alpha$  can be represented by

$$G(\delta, \alpha) = S(\alpha) G(\delta) S(-\alpha) \quad (37)$$

A series of optical elements can be represented by the product of a series of matrices. This simplifies enormously the task of computing the effect of many elements. It is also possible with the Jones calculus to derive a series of general theorems concerning combinations of optical elements. Jones has described three of these, all of which apply only for monochromatic light.

1. An optical system consisting of any number of retardation plates and rotators is optically equivalent to a system containing only two elements, a retardation plate and a rotator.

2. An optical system containing any number of partial polarizers and rotators is optically equivalent to a system containing only two elements—one a partial polarizer and the other a rotator.

3. An optical system containing any number of retardation plates, partial polarizers, and rotators is optically equivalent to a system containing four elements—two retardation plates, one partial polarizer, and one rotator.

As an example of the power of the calculus, a rather specific theorem can be proved. A rotator of any given angle  $\alpha$  can be formed by a sequence of three retardation plates, a quarter-wave plate, a retardation plate at  $45^\circ$  to the quarter-wave plate, and a second quarter-wave plate crossed with the first.

$$S(\alpha) = S(\beta)G\left(-\frac{\pi}{2}\right)S(-\beta)S\left(\beta + \frac{\pi}{4}\right)G(\delta)S\left(-\beta - \frac{\pi}{4}\right)S(\beta)G\left(\frac{\pi}{2}\right)S(-\beta) \quad (38)$$

where  $\beta$  is the angle between the axis of the first quarter-wave plate and the  $x$  axis, and  $\delta$  is the retardation of the plate in the middle of the sandwich.

The first simplification arises from the fact that the axis rotations can be done in any order.

$$S\left(\beta + \frac{\pi}{4}\right) = S(\beta)S\left(\frac{\pi}{4}\right) = S\left(\frac{\pi}{4}\right)S(\beta) \quad (39)$$

This reduces the equation to

$$S(\alpha) = S(\beta)G\left(-\frac{\pi}{2}\right)S\left(\frac{\pi}{4}\right)G(\delta)S\left(-\frac{\pi}{4}\right)G\left(\frac{\pi}{2}\right)S(-\beta) \quad (40)$$

$$\text{Now } S\left(-\frac{\pi}{4}\right)G\left(\frac{\pi}{2}\right) = \frac{1}{\sqrt{2}} \begin{vmatrix} 1 & -i \\ -i & 1 \end{vmatrix} \quad (41)$$

$$\text{and } G\left(-\frac{\pi}{2}\right)S\left(\frac{\pi}{4}\right) = -\frac{1}{\sqrt{2}} \begin{vmatrix} 1 & i \\ i & 1 \end{vmatrix} \quad (42)$$

When the multiplication is carried through

$$S(\alpha) = S(\beta) - \frac{1}{2} \begin{vmatrix} e^{i\delta/2} + e^{-i\delta/2} & -ie^{i\delta/2} + ie^{-i\delta/2} \\ ie^{i\delta/2} - ie^{-i\delta/2} & e^{i\delta/2} + e^{-i\delta/2} \end{vmatrix} S(-\beta) \quad (43)$$

$$= S(\beta)S\left(\frac{\delta}{2}\right)S(-\beta) = S\left(\frac{\delta}{2}\right) \quad (44)$$

The rotation angle is therefore equal to one-half the phase angle of the retardation. This combination is a true rotator in that the rotation is independent of the azimuth angle of the incident polarized light.

A variable rotator can be made by using a Soleil compensator for the central element. This consists of two quartz wedges joined to form a plane parallel plate. The lower wedge is cemented to a plane parallel quartz plate.

**Mueller matrices.** In the Jones calculus, the intensity of the light passing through the system must be obtained by calculation from the components of the light vector. A second calculus is frequently used in which the light vector is split into four components. This also uses matrix operators which are termed Mueller matrices. In this calculus, the intensity  $I$  of the light is one component of the vector and thus is automatically calculated. The other components of the vector are

$$M = A_x^2 - A_y^2 \quad (45)$$

$$C = 2A_x A_y \cos(\varphi_x - \varphi_y) \quad (46)$$

$$S = 2A_x A_y \sin(\varphi_x - \varphi_y) \quad (47)$$

The matrix of a perfect polarizer parallel to the  $x$  axis can be written

$$P = \frac{1}{2} \begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix} \quad (48)$$

This calculus can treat unpolarized light directly. Such a light vector is given by the expression

$$\begin{vmatrix} I \\ M \\ C \\ S \end{vmatrix} = \begin{vmatrix} 1 \\ 0 \\ 0 \\ 0 \end{vmatrix} \quad (49)$$

The vector for light polarized parallel to the  $x$  axis is written

$$\begin{vmatrix} 1 \\ 1 \\ 0 \\ 0 \end{vmatrix} \quad (50)$$

In the same manner as in the Jones calculus, matrices can be derived for retardation plates, rotators, and partial polarizers. This calculus can also be used to derive various general theorems about various optical systems. See FARADAY EFFECT; INTERFERENCE OF WAVES; OPTICAL ACTIVITY; ROTATORY DISPERSION.

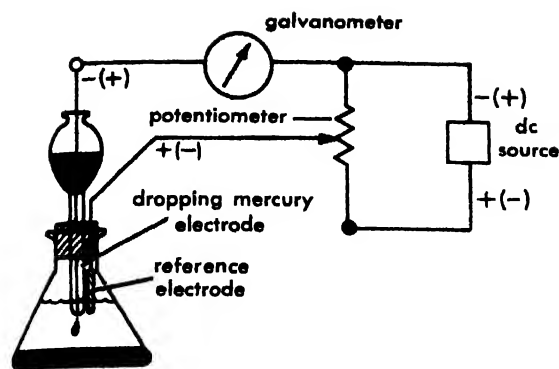
[B. H. BILLINGS]

**Bibliography:** M. Born and E. Wolf, *Principles of Optics*, 1959; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 3d ed., 1957; R. C. Jones, The Sohncke theory of optical activity, *J. Opt. Soc. Am.*, 31(7):488-503, 1941; J. Strong, *Concepts of Classical Optics*, 1958.

## Polarographic analysis

An electrochemical technique used in analytical chemistry. Polarography is one of a group of techniques which are broadly classed as voltammetric. The common feature of these techniques is that they all involve observation of current-potential time relationships at electrodes immersed in electrolytic solutions. The technique can be applied to any ionic or molecular species that is oxidizable or reducible in electrolysis within certain limits of potential. Most metals, and some organic functional groups such as aldehyde, ketone, amino, and mercaptan may be determined by polarography.

In conventional polarography, the current through the electrolysis cell is measured as a function of the applied potential. At sufficiently anodic (oxidizing) potentials, no current is observed in the case of a reducible substance. As the potential becomes more cathodic, appreciable reduction occurs and the current rises in approximately exponential fashion. Ultimately, a potential is reached at which reduction occurs as rapidly as the ions or molecules can reach the electrode surface. When this happens, the current levels off at a limiting value independent of further increase in potential. Thus the current-potential curve or wave has a sigmoidal shape. The magnitude of the limiting current is proportional to the concentration



Polarograph.

tion of the reducible substance so the technique has value in quantitative analysis. The half-wave potential, the potential at which the current is midway between the residual current (the current not attributable to the reducible substance in question) and limiting current plateau, is often within a few millivolts of the standard potential for the electrode reaction. Even when it is not, the half-wave potential is still characteristic of the reacting species so that qualitative analysis is also possible. Oxidations are studied in an analogous way.

**Apparatus.** The system consists of a potentiometer for adjusting the potential, a galvanometer for measuring current, and a cell. The cell commonly contains two electrodes, a reference electrode (either calomel or silver-silver chloride), the potential of which is constant, and an indicator electrode.

The most widely used polarographic indicator electrode is the dropping-mercury electrode, which consists of a fine-bore capillary tube above which a constant head of mercury is maintained. The mercury emerges from the tip of the capillary at the rate of a few milligrams per second. It forms spherical droplets which fall from the capillary orifice into the solution at the rate of one every 2–10 sec.

This novel electrode has a number of great advantages over other possible arrangements. First, mercury has a high hydrogen overvoltage so that processes which would be obscured by decomposition of water at other electrodes can be readily observed at its surface. Second, the fact that the electrode surface is renewed periodically means that complications due to change in surface composition during the process are not encountered. The solution in contact with these electrodes is  $10^{-2}$  to  $10^{-5}$  molar in the electroactive species of ion or molecule and contains, in addition, a large excess (fiftyfold or greater) of a supporting electrolyte, that is, one which does not react at the electrode in the potential region of interest. The function of the latter is to reduce the resistance of the solution and thus to ensure that diffusion in a concentration gradient rather than migration in an ohmic potential gradient is the mode of transfer of the species to the electrode surface.

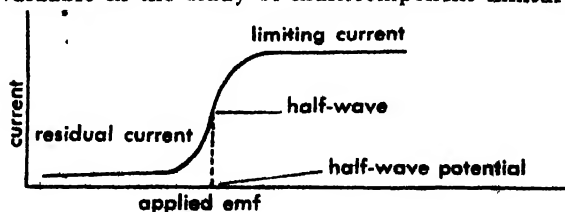
**Applications.** The limitations of the potential range are set on the cathodic side by the reduction of the solvent or supporting electrolyte and on the anodic side by oxidation of these or of the electrode material.

The method has been used for the determination of most metals, including those as active as sodium and potassium, and for the determination of various organic functional groups including aldehydes, ketones, halogens, double and triple bonds, nitro, nitroso, azo, hydrazo, amino, sulfides, mercaptans, and many others. Because mercury ions react with various anions, anodic dissolution currents of mercury can be used for the determination of halides, thiocyanate, cyanide, and sulfide among others. Oxygen is also reduced at the dropping-mercury electrode. Although this forms the basis for a convenient oxygen determination, it means that solutions to be analyzed for other constituents must be purged of oxygen.

A polarogram can be plotted manually by varying the potential in finite increments and measuring the current at each. Alternatively, commercial instruments are available in which the potential is scanned by a motor-driven slide wire and the current is recorded automatically. The observed currents are usually a few microamperes. The accuracy of a quantitative analysis is about 1%.

**Other polarographic techniques.** Various electrodes other than the dropping-mercury electrode have been used, including dropping, rotating, vibrating, or stationary electrodes of tungsten, gold, silver, platinum, gallium, graphite, mercury or mercury amalgam pools or stationary drops, and mercury-plated platinum and gold. Although solid electrodes have the disadvantage of giving erratic results, they can be used at much more anodic potentials than mercury.

In addition to the conventional potential scan method, other techniques based on the same principles have been applied in special situations. So-called oscillographic polarography in which alternating-current (ac) potentials are applied to the electrode, or in which a conventional potential scan is performed in 1 millisecond or less, is particularly useful where highest sensitivity is desired or where the kinetics of fast reactions are studied. Conventional scans, in which a small (a few millivolts) sinusoidal or square-wave potential is superimposed on the direct-current (dc) potential and the ac component of the resulting current is measured, give curves which are essentially derivatives of the normal polarographic wave. These techniques are valuable in the study of multicomponent mixtures



Polarographic curve.



where the waves of the separate components are not sufficiently separated for conventional analysis. Inverse methods in which species are plated into mercury or onto the surfaces of solid electrodes and their dissolution currents measured have been applied to analysis of solutions as dilute as one part per  $10^{12}$  parts of solution.

Many of these techniques are also useful for following the course of titrations. With a constant potential applied, the amperometric titration curve of current as a function of titrant added takes the form of two straight-line segments with their intersection at the stoichiometric point.

**Other voltammetric methods.** There are other voltammetric techniques in which current is controlled and potential is measured. If a linearly changing current is applied to a DME or rotated solid electrode, a curve of the same form as a conventional polarogram is obtained. This is current-scan polarography. A technique of more recent importance is chronopotentiometry. In this technique, a constant current is applied to a stationary electrode in unstirred solution and the potential is measured as a function of time. The time at which a large potential break is observed is termed the transition time, and its square root is proportional to the concentration of the reacting species. If the data are plotted as square root of time versus potential, the resulting curve is of the same form as a conventional polarogram. See OVERVOLTAGE; TITRATION, AMPEROMETRIC. [W. H. REINMUTH]

**Bibliography:** I. M. Kolthoff and J. J. Lingane, *Polarography*, 2d ed., 1952; J. J. Lingane, *Electroanalytical Chemistry*, 2d ed., 1958.

## **Poliomyelitis**

An acute infectious viral disease which in its serious form affects the central nervous system and, by destruction of motor neurons in the spinal cord, produces flaccid paralysis. However, about 99% of infections are either inapparent or very mild. See CENTRAL NERVOUS SYSTEM.

**Infectious agent.** Polioviruses are no longer considered strictly neurotropic, because (1) they will multiply in cultures of many nonnervous tissues; (2) viremia and antibody formation appear before the paralytic phase is reached and in cases where even transient signs of central nervous system involvement do not occur; (3) the virus is regularly found in the throat and stools before the onset of disease, and after onset it is found for a week in the throat, and for several weeks in the stools. The infective virus particle is about 28 m $\mu$  in diameter. Freezing preserves it for long periods. In contrast to the arbor viruses, it is destroyed only slowly by alcohol and not at all by ether and deoxycholate. Poliovirus has a very restricted host range; most strains are limited in vivo to primates and in vitro to primate tissue cultures: monkey kidney cultures are widely used. Three antigenic types of poliovirus are known, but most clinical cases are due to Type 1. See ANIMAL VIRUS; ARBOR VIRAL ENCEPHALITIDES; ENTEROVIRUS.

**Pathogenesis.** The virus probably enters the body through the mouth; primary multiplication occurs in the throat and intestines. Transitory viremia occurs; the blood seems to be the most likely route to the central nervous system. The severity of the infection may range from a completely inapparent through minor influenzalike illness, or an aseptic meningitis syndrome (nonparalytic poliomyelitis) with stiff and painful back and neck, to the severe forms of paralytic and bulbar poliomyelitis. In all clinical types, virus is regularly present in the enteric tract. In paralytic poliomyelitis the usual course begins as a minor illness but progresses, sometimes with an intervening recession of symptoms (hence biphasic), to flaccid paralysis of varying degree and persistence. When the motor neurons affected are those of the diaphragm or of the intercostal muscles, respiratory paralysis occurs. Bulbar poliomyelitis results from viral attack on the medulla (bulb of the brain) or higher brain centers, with respiratory, vasomotor, facial, palatal, or pharyngeal disturbances.

**Diagnosis.** Laboratory diagnosis is by isolation, usually from stools inoculated into tissue cultures, and subsequent identification by neutralization with specific antisera in vitro, and by complement-fixing and neutralizing serum antibody rises. Isolation of a virus which is cytopathogenic in tissue cultures is not sufficient for diagnosis, for many other cytopathogenic enteroviruses also inhabit the enteric tract and produce syndromes similar to mild or early poliomyelitis. See CULTURE, TISSUE; NEUTRALIZING ANTIBODY.

**Epidemiology.** Poliomyelitis occurs throughout the world. In temperate zones it appears chiefly in summer and fall, although winter outbreaks have been known. It occurs in all age groups, but less frequently in adults because of their acquired immunity. In crowded underdeveloped areas and in tropical countries, where conditions favor a constant widespread dissemination of virus, poliomyelitis continues to be a disease of infancy, and a high percentage of children over 4 years old are already immune. During recent decades in some areas of the temperate zones the age incidence has tended to change in marked parallel with improving sanitation and hygiene; exposure is postponed sometimes so long that parents may be without immunity and succumb to infections transmitted from their children. Even in epidemic periods before a vaccine was available, the proportion of poliomyelitis infections resulting in paralysis or meningitis was relatively small. In 1952, the year of maximum incidence in the United States, one case occurred in every 3000 persons of all ages, and one in every 1000 children.

The virus is spread by human contact; the nature of the contact is not clear, but it appears to be associated with familial contact and with interfamily contact among young children. The virus may be present in flies.

**Vaccine.** Salk vaccine (formalin-killed), prepared from virus grown in monkey kidney cultures,

has been widely used in the United States. Oral poliovirus vaccine has also been introduced and is now accepted throughout the world as an effective immunizing agent. The vaccine is a living, attenuated virus, and like most viruses it is unstable except when held at very low temperatures in the frozen state. The use of magnesium chloride as a stabilizing agent for poliovirus overcomes this problem for in molar concentrations  $MgCl_2$  protects live poliovirus vaccines, so that they may be stored in the ordinary refrigerator for over a year with no loss in immunizing potency.  $MgCl_2$ -stabilized vaccines also maintain potency over periods of several weeks at room or transit temperatures. Thus, the stabilized vaccine offers numerous advantages not only in the physician's office but also in mass poliovirus immunization programs in the field, particularly in underdeveloped and tropical countries, where it is difficult to maintain or transport vaccine under frozen conditions.

[J. L. MLINICK]

**Bibliography:** International Poliomyelitis Congress, *Poliomyelitis; papers and discussions presented at the Fifth International Poliomyelitis Conference, 1961.*

## Polishing

The smoothing of a surface by the cutting action of abrasive grit either glued to or impregnated in a flexible wheel or belt. Polishing, not a precision process, removes stock until the desired surface condition is obtained. Wheels are built up from layers of soft materials, such as wood or leather.

When a considerable amount of stock must be removed, operations may start with a coarse grit but use a finer grit wheel for finishing. Common polishing machines may be either bench or floor mounted. Usually they are lathe type machines with wheels at either end of a power spindle, or one end may drive an abrasive belt. Various types of semiautomatic polishing machines are used for quantity production. These are designed to move the workpieces through the cutting paths of one or more polishing wheels. See GRINDING.

For a mirrorlike surface, a precision abrading process called superfinishing is used. Superfinishing removes minute flaws or inequalities. While it is not primarily intended to remove stock, generally a dimensional change of .0001-.0002 in. on diameter occurs. Superfinishing is performed with an extremely fine grit abrasive stone, shaped to match and cover a large portion of the work surface. As the workpiece turns, the stone reciprocates across the work under a flood of lubrication. The process is usually performed on symmetrical pieces. See LAPPING; MACHINING OPERATIONS.

[A. TUTTLE]

## Pollucite

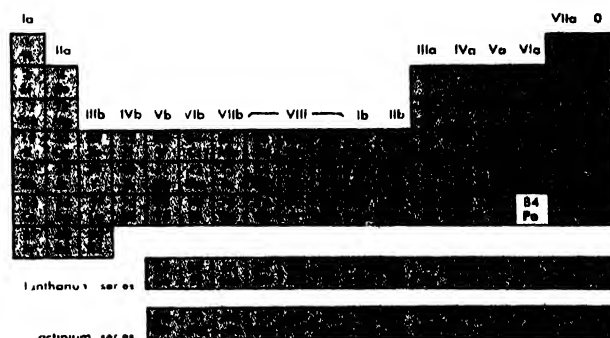
A mineral tectosilicate with composition  $Cs_3Al_4Si_9O_{26} \cdot H_2O$ . Pollucite crystallizes in the isometric system but well-formed crystals, which are usually cubic, are rare; it is most commonly massive. The

hardness is  $6\frac{1}{2}$  on Mohs scale, and the specific gravity is 2.90. The mineral is colorless to white with a vitreous luster. Pollucite is a rare mineral found in pegmatites associated with spodumene, lepidolite, eucryptite, petalite, and cesium beryl. It has been found in large masses at the Varuträsk pegmatite in Sweden. In the United States it was at one time worked as an ore of cesium at Newry, Maine, and in the Black Hills of South Dakota. See SILICATE MINERALS.

[C. S. HURLBUT, JR.]

## Polonium

A chemical element, Po, atomic number 84, a member of group VIa of the periodic table. Marie Curie discovered the radioisotope  $Po^{210}$  in pitchblende and named it after her native Poland. This isotope, also known as radium F, is the penultimate member of the radium decay series; pitchblende contains 0.1 mg/ton. All polonium isotopes are radioactive, and all are short-lived except the three  $\alpha$ -emitters:  $Po^{208}$  (2.9 years),  $Po^{209}$  (100 years), both of which are produced by bombarding bismuth



with deuterons, and natural  $Po^{210}$  (138.4 days), now produced in milligram amounts by the neutron bombardment of bismuth. Polonium is separated from bismuth by spontaneous deposition onto a less noble metal, such as silver, followed by a vacuum sublimation or chemical separation of the deposit.

**Uses.** Polonium ( $Po^{210}$ ) is used mainly for the production of neutron sources; for these, the polonium is alloyed with elements such as beryllium which have isotopes of high  $\alpha, n$  cross section. It can also be used in static eliminators and, when incorporated in the electrode alloy of spark plugs, is said to improve the cold-starting properties of internal combustion engines.

**Properties.** Most of the chemistry of polonium has been determined using  $Po^{210}$ , 1 curie of which weighs 222.2  $\mu g$ ; work with weighable amounts is therefore hazardous and requires special techniques.

Polonium is more metallic than its lower homolog, tellurium; two allotropes are known,  $\alpha$ -Po (simple cubic) and  $\beta$ -Po (simple rhombohedral) with the phase change  $\alpha \rightarrow \beta$  at about  $36^\circ C$ . The metal is chemically similar to tellurium, forming the bright red compounds  $SPoO_3$  and  $SePoO_3$ ; a number of polonides are known.

The metal is soft, and its physical properties resemble those of thallium, lead, and bismuth. Valences of 2 and 4 are well established, and there is some evidence of hexavalency. Polonium lies between silver and tellurium in the electrochemical series.

**Compounds.** Two forms of the dioxide are known: low-temperature, yellow, face-centered cubic (UO<sub>2</sub> type), and high-temperature, red, tetragonal. It is formed from the elements at 250°C and decomposes at 500°C under vacuum. The black monoxide may be formed in the spontaneous decomposition of SPoO<sub>3</sub> and SePoO<sub>3</sub>. The quadrivalent hydroxide (pale yellow, gelatinous, feebly amphoteric) is precipitated from solutions of polonium salts by alkalis; it is reduced to the metal in alkaline suspension by hydroxylamine, hydrazine, sodium dithionite, and ammonia (liquid or concentrated aqueous solution). The last two reactions may be due to  $\alpha$ -radiation effects. The brown, bivalent hydroxide is readily oxidized. Polonium monosulfide (black) is precipitated from acid solutions of polonium salts by hydrogen sulfide (solubility product  $5.5 \times 10^{-29}$ ); it decomposes to the elements at 275°C and 10 $\mu$  pressure, a property utilized in the preparation of pure polonium metal.

The halides are covalent, volatile compounds, resembling their tellurium analogs (PoCl<sub>4</sub>, yellow; PoBr<sub>4</sub>, carmine; PoI<sub>4</sub>, black; PoCl<sub>2</sub>, ruby-red; PoBr<sub>2</sub>, purple-brown; PoCl<sub>2</sub>Br<sub>2</sub>, salmon-pink) with the bivalent state more stable than the quadrivalent; the latter gives rise to face-centered cubic complex salts M<sub>2</sub>PoX<sub>6</sub> (X = Cl, yellow; Br, brick-red; I, black), where M is a univalent cation. The cesium salts are the least soluble. The quadrivalent halides in solution are reduced to the bivalent state (pink) by hydrazine, sulfur dioxide, or arsenic(III) oxide (on warming) and to the metal by tin(II) chloride, sodium dithionite, or titanium(III) chloride. Complexes with organic molecules (for example, tributyl phosphate, dithizone, ethylenediaminetetraacetate), nitrosyl chloride, and ammonia are also known.

The increased metallic character of polonium as compared to tellurium is shown in the formation of the salts Po(SO<sub>4</sub>)<sub>2</sub> (white, hydrated; deep purple, anhydrous) and Po(NO<sub>3</sub>)<sub>4</sub> (white). Both are readily hydrolyzed, the former to 2PoO<sub>2</sub>·SO<sub>3</sub> (white when cold and yellow when hot), the latter to a series of basic nitrates which may be polymeric. The basic salts 2PoO<sub>2</sub>·SeO<sub>3</sub> and 2PoO<sub>2</sub>·CrO<sub>3</sub> are also known; they are analogous to tellurium sulfate, 2TeO<sub>2</sub>·SO<sub>3</sub>. There is some evidence for a normal chromate (yellow Po(CrO<sub>4</sub>)<sub>2</sub>?) and a hexavalent polonium/chromium complex acid. The acetate, cyanide, iodate, oxalate, and phosphate (all white) have been prepared, and there is evidence for the formation of a vanadate and tartrate.

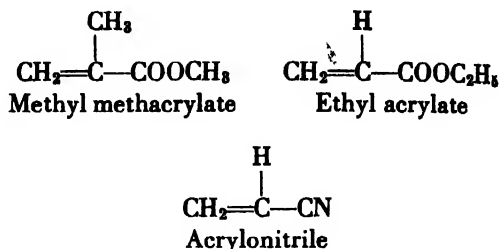
**Determination.** Polonium (Po<sup>210</sup>) is estimated by its  $\alpha$ -emission, either by direct counting or calorimetrically; in the latter case, the measurement depends on the heat liberated by the stoppage of the disintegration  $\alpha$ -particles within the

sample. See NUCLEAR REACTION; RADIOACTIVITY; TELLURIUM.

**Bibliography:** K. W. Bagnall, *Chemistry of the Rare Radioelements*, 1957; K. W. Bagnall et al., *The Polonium Chemistry Project*, Atomic Energy Research Establ. (Gt. Brit.), C/R 2566, 1958.

## Polyacrylate resin

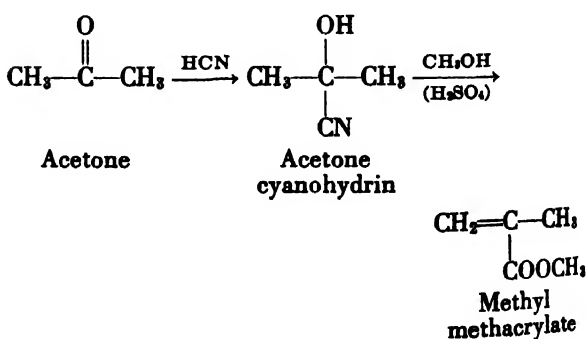
Useful polymers can be obtained from a variety of acrylic monomers, such as acrylic and methacrylic acids, their salts, esters, and amides, and the corresponding nitriles. The most important monomers are



Polymethyl methacrylate, ethyl acrylate, and a few other derivatives are discussed below. See POLYACRYLONITRILE RESIN.

Polymethyl methacrylate is distinguished as a hard, transparent polymer with high optical clarity, high refractive index, and good resistance to light and aging. It and its copolymers are useful for lenses, signs, indirect lighting fixtures, transparent domes and skylights, dentures, and protective coatings.

The monomer may be prepared by the dehydration and methanolysis of acetone cyanohydrin:



Polymerization may be initiated by free-radical catalysts, such as peroxides, or by organometallic compounds, such as butyl lithium. The free-radical polymerization can be carried out in bulk, in solution, and in aqueous emulsion or suspension. Although solution polymerization is not commonly used, bulk polymerization is frequently employed in various casting operations, as in the formation of sheets, rods, and tubes, in the mounting of biological, textile, and metallurgical test specimens, and in dental applications. A 20% reduction in volume accompanies the conversion of monomer to polymer. This makes it difficult to prepare articles to predetermined dimensions by the cast polymeri-

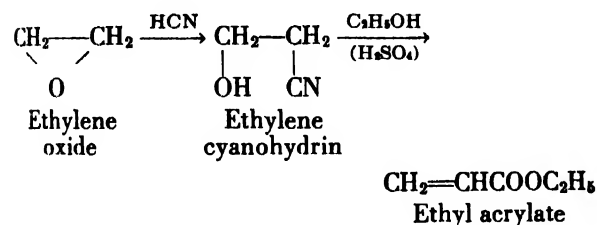
zation technique. The problem is largely minimized by using a syrup of polymer dissolved in monomer or a dough of finely divided polymer dispersed in a relatively small amount of monomer.

Molding powders suitable for the injection molding of dials, ornamental fixtures, and lenses may be prepared from the granules produced by aqueous suspension polymerization or from the product of bulk polymerization.

Solutions of polymethyl methacrylate and its copolymers are useful as lacquers. Aqueous latexes formed by the emulsion polymerization of methyl methacrylate with other monomers are useful as water-based paints and in the treating of textiles and leather.

Polyethyl acrylate is a tough, somewhat rubbery product. The monomer is used mainly as the plasticizing or softening ingredient in copolymers. The relatively hard or leathery polymers of methyl methacrylate, vinyl acetate, or vinyl chloride can be made softer by adding moderate amounts of ethyl acrylate in the original polymerization.

Ethyl acrylate may be produced by the dehydration and ethanolysis of ethylene cyanohydrin which can be obtained from ethylene oxide:



The monomer is also produced by the reaction of acetylene, carbon monoxide, and ethyl alcohol in the presence of nickel carbonyl.

Polymerization may be effected by catalysts of the free-radical type. Copolymerizations with other monomers are frequently carried out in aqueous emulsion or suspension.

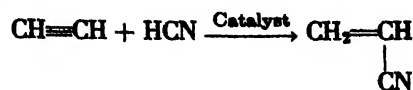
Other acrylic derivatives have been the subject of continuing interest. Methyl chloroacrylate and cyclohexyl methacrylate yield polymers with relatively high softening points and resistance to scratching. The butyl and octyl esters of acrylic acid yield rubbery materials. Addition of polylauryl methacrylate to petroleum lubricating oil improves the flowing properties of the oil at low temperatures and the resistance to thinning at high temperatures. See ACRYLONITRILE; PLASTICS FABRICATION; POLYMERIZATION. [J.A.M.; L.M.H.]

## Polyacrylonitrile resin

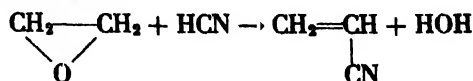
A hard, horny, relatively insoluble, and high-melting material. Polyacrylonitrile (polyvinyl cyanide) is used almost entirely in copolymers. The copolymers fall into three groups: fibers, plastics, and rubber. The presence of acrylonitrile in a polymeric composition tends to increase its resistance to temperature, chemicals, impact, and flexing.

Acrylonitrile is generally prepared by one of the following methods:

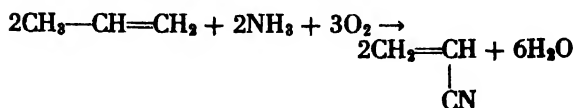
(1) Catalyzed addition of hydrogen cyanide to acetylene



(2) Reaction between hydrogen cyanide and ethylene oxide



(3) Reaction between ammonia and propylene



The polymerization of acrylonitrile can be readily initiated by means of the conventional free-radical catalysts, such as peroxides, by irradiation, or by the use of alkali metal catalysts. Although polymerization in bulk proceeds too rapidly to be commercially feasible, satisfactory control of a polymerization or copolymerization may be achieved in suspension and in emulsion, and in aqueous solutions from which the polymer precipitates. Copolymers containing acrylonitrile may be fabricated in the manner of thermoplastic resins.

The major use of acrylonitrile is in the form of fibers. The high strength, high softening temperature, resistance to aging, chemicals, water, and cleaning solvents, and the soft wool-like feel of fabrics have made the product popular for many uses such as sails, cordage, blankets, and various types of clothing. Commercial forms of the fiber probably are copolymers containing minor amounts of other vinyl derivatives, such as vinyl pyrrolidone, vinyl acetate, vinyl chloroacetate, acrylamide, or others. The comonomers are included to produce specific effects, such as improvement of dyeing qualities.

Extensive use is made of copolymers of acrylonitrile with butadiene, often called Buna N rubbers, which contain 15–40% acrylonitrile. Minor amounts of other unsaturated esters, such as ethyl acrylate, which yield carboxyl groups on hydrolysis may be incorporated to improve the curing properties. The Buna N rubbers resist hydrocarbon solvents, such as gasoline, and abrasion and in some cases show high flexibility at low temperatures.

In recent years, blends of vinyl polymers, such as polystyrene or polyvinyl chloride with small to moderate amounts of acrylonitrile or with an acrylonitrile-butadiene copolymer, have represented a significant advance in polymer technology. The products combine the hardness of the vinyl polymer and the impact resistance of the rubbery components. Some of the many applications are as molding compounds for products possessing high impact resistance, such as pipe, and as sheets for structural uses, such as industrial ducts, refrigerator liners, and orthopedic devices. See ACRYLONITRILE.

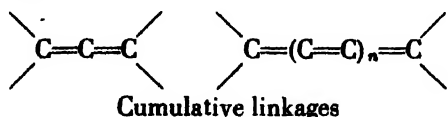
## Polyalkene

TRILE; PLASTICS FABRICATION; POLYMER PROPERTIES; RUBBER. [J.A.M.; L.M.H.]

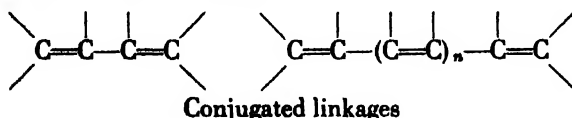
### Polyalkene

One of a class of organic compounds containing two or more ethylenic linkages in the molecule. These compounds are sometimes termed polyenes. They exist in the following three systems:

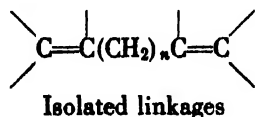
1. The unsaturated linkages may be directly attached. These compounds are said to possess cumulative unsaturation.



2. The unsaturated linkages may alternate with single linkages, in which case the unsaturation is said to be conjugated.



3. The unsaturated linkages may be separated by one or more carbons, in which case the unsaturation is said to be isolated.



Allene compounds are mainly of interest in studies involving the stereochemistry of organic compounds, since it is possible to synthesize optically active compounds which are mirror-images of each other. See OPTICAL ACTIVITY.

Conjugated dienes are the most important group of polyenes because such compounds as butadiene, isoprene, and cyclopentadiene are included in this classification. See DIENE.

Isolated polyenes include the unsaturated hydrocarbon squalene which contains six isoprene units with six isolated double bonds. It is an aliphatic triterpene and is related to the carotenoids such as lycopene, the red coloring matter in tomatoes, and carotene, pro-vitamin A. See ALKENE; TERPENE; TRITERPENE. [C.A.C.]

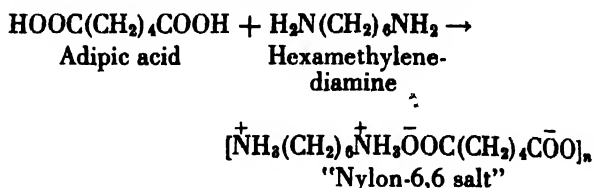
**Bibliography:** H. Gilman (ed.), *Organic Chemistry*, vol. 1, 2d ed., 1943.

### Polyamide resin

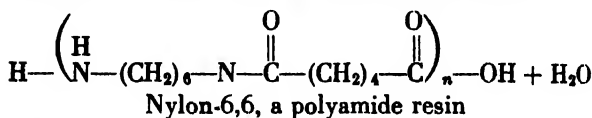
A horny, whitish, translucent, high-melting polymer. Polyamide resins can be essentially transparent and amorphous when their melts are quenched. On annealing or cold drawing, they become highly crystalline and translucent. The polymers are used for fibers, bristles, bearings, gears, molded objects, coatings, and adhesives. The term nylon refers specifically to synthetic polyamides which are capable of forming fibers. See FIBER, MAN-MADE.

Brief outlines of the preparations of commercial polyamides by (1) the reaction of dicarboxylic acids with diamines, (2) the condensation of amino acids, and (3) the reaction of so-called polymerized vegetable-oil acids with polyamines are given in this article.

Nylon-6,6 and nylon-6,10 are products of the condensation reaction of hexamethylenediamine (6 carbon atoms) with adipic acid (6 carbon atoms), and with sebacic acid (10 carbon atoms), respectively. By heating equimolar proportions of the two reactants, a polymeric salt is formed,



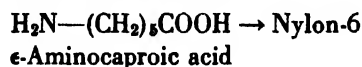
which, on further heating, yields the polyamide



Because the end groups on the polymer can react on further heating, as in melt spinning, it is desirable to add a very small amount of a monoacid or monoamine to the polymerizing mixture in order to prevent the formation of material of very high molecular weight.

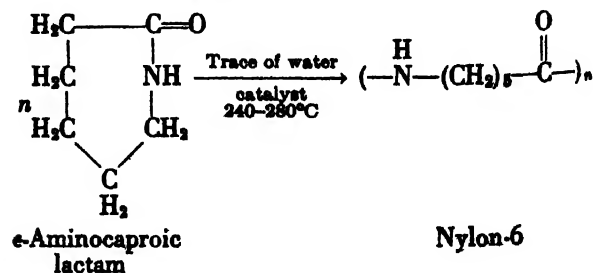
Nylon-6,6 and nylon-6,10 are used primarily for fibers; however, they are also employed in molding compositions.

Nylon-6 and nylon-11 are obtained by the self-condensation of  $\epsilon$ -aminocaproic acid and  $\omega$ -aminoundecanoic acids, respectively,



Each molecule, containing both the amino and carboxylic groups, can condense to yield high polymers by reactions similar to those between the diacids and the diamines.

Nylon-6 can also be prepared by the polymerization of the lactam of  $\epsilon$ -aminocaproic acid,



Nylon-6 and -11 are used mainly as molding compounds for the production of machine parts, such

as gears and bearings, and for electrical mountings; they are also used as filaments, bristles, and films. Nylon bearings and gears perform quietly and need little or no lubrication.

The unsaturated fatty acids in vegetable oil (for example, linoleic acid) may dimerize or polymerize to low polymers through their unsaturated groups. The di- or polycarboxylic acids obtained yield tough polyamides by condensation with di- or polyamines. These products are employed as coatings and adhesives and are also used in epoxy and phenolic resin formulations. See PLASTICS FABRICATION. POLYMERIZATION. [J.A.M.; L.M.H.]

## Polychaeta

A class of the Annelida sometimes called Chaetopoda. These organisms are largely free-living, littoral in neritic (coastal or shoreline) localities or planktonic in the open sea, or in deep ocean bottoms. They are found in all parts of the world and occur at all depths, but are best known from the upper 150 fathoms.

**Taxonomy.** Polychaetes are classified into 64 families containing about 1600 genera and 10,000 species. They vary in size according to the species. The largest are in the superfamily Eunicia and measure to 900 mm in length. The smallest, among the Archannelida, measure less than 1 mm. The extent of their diversity is exemplified in common or generic names such as in the following list.

Common name	Genus
Sea mouse	<i>Amphitrite</i> ; <i>Aphrodite</i>
Loose worm	<i>Arenicola</i>
Fringe worm	<i>Cirratulus</i>
Proboscis or bloodworm	<i>Eunice</i> ; <i>Glycera</i>
Clim worm	<i>Neanthes</i> , <i>Nereis</i>
Sand worm	<i>Nephtys</i>
Mud blister worm	<i>Polydora</i>
Feather duster worm	<i>Sabella</i> ; <i>Serpula</i>
Golden crown worm	<i>Pectinaria</i>
Gooseberry worm	<i>Sternopsis</i>
Reef builder worm	<i>Sabellaria</i>

Other descriptive names include fireworms (the Amphinomidae), because of the stinging sensation caused when their harpoon-setae penetrate the skin, and the bamboo worms (the Maldanidae).

The large group of Polychaeta are conveniently divided into the Errantia, meaning freely moving, and the Sedentaria, meaning tubicolous, but these terms have no strict morphological application because many of the first are tubicolous and some of the second are errant. The use of the terms Phanerocephala, meaning head exposed, and Cryptoccephala, or head concealed, parallels that of the first set of words. It is impossible to separate the vast assemblage of families except to ally certain groups, such as the Aphroditea for the scale-bearing worms, the Eunicia for those with characteristic jaws, and the Serpulea for those with a pinately divided crown. This leaves the majority of families and species unattached. It is customary to

place the elytral-bearing aphroditids at one end, and the operculated serpulids at the other. The chart indicates the arrangement as currently used. Errantia

### Scale worms and allied families

Aphroditidae  
Polynoidae  
Polyodontidae  
Sigalionidae  
Pareulepidae  
Chrysopetalidae  
Palmyridae  
Pisionidae

### Fireworms and allied families

Amphinomidae  
Euprosinidae  
Spintheridae

### Leaf worms and allied pelagic families

Phyllodoceidae  
Lopadorhynchidae  
Iospilidae  
Pontodoridae  
Lacydonidae  
Alciopidae  
Typhloscolecidae  
Tomopteridae (Gymnocopa)  
Hesionidae  
Palaigiidae  
Syllidae  
Nereidae  
Nephtyidae  
Sphaerodoridae  
Glyceridae  
Goniadidae

### Superfamily Eunicia

Onuphidae  
Eunicidae  
Lumbrineridae  
Arabellidae  
Lysaretidae  
Dorvilleidae

### Parasitic families

Histiobdellidae  
Ichtyotomidae  
Myzostomidae  
Myzostomidae  
Protomyzostomidae  
Mesomyzostomidae  
Stelechopidae

### Sedentaria

Orbiniidae  
Paraonidae  
Aristobranchidae

### Anterior end with a pair of long palpi

Spionidae  
Magelonidae  
Longosomidae  
Disomidae  
Chaetopteridae

### Fringe worms

Cirratulidae  
Ctenodrilidae



### Limivores

Scalibregmidae  
Opheliidae  
Sternaspidae  
Capitellidae  
Arenicolidae  
Maldanidae  
Oweniidae

### Anterior end with retractile oral tentacles

Flabelligeridae  
Sabellariidae  
Pectinariidae (cone worms)  
Terebellidae  
Trichobranchidae

### Feather duster worms

Sabellidae  
Serpulidae

### MORPHOLOGY

**Head.** The head or prostomium may be a simple or secondarily annulated lobe in front of, or above, the mouth; it may be pushed far back (Fig. 1), or be retractile into one of the first few segments. It may have simple eyespots or complex, lenticular eyes. Complex eyes are best developed in pelagic polychaetes. Antennae are sensory, threadlike structures (Fig. 2a,b) in various arrangements. A caruncle or fleshy sensory organ is characteristic of many Errantia and is sometimes highly developed, as in the fireworms, *Hermodice* (Fig. 3) and *Amphinome* (Fig. 2a).

**Peristomium.** The peristomium or first ring may be a simple smooth ring surrounding the mouth with its ventral part forming the lower lip, or it may have setae and be variously modified to form complex structures. Examples are the tentacular crown (Fig. 4) of a sabellid, *Fabricia*, and the

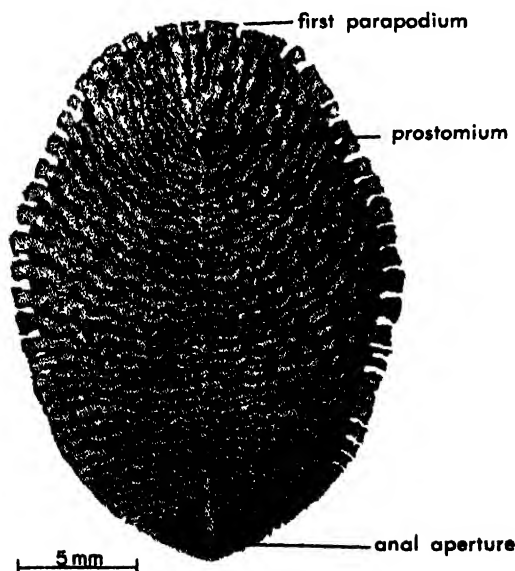


Fig. 1. *Spinther*, in dorsal view, showing the small prostomium set far back, the dorsal anus, and the transverse parapodial folds.

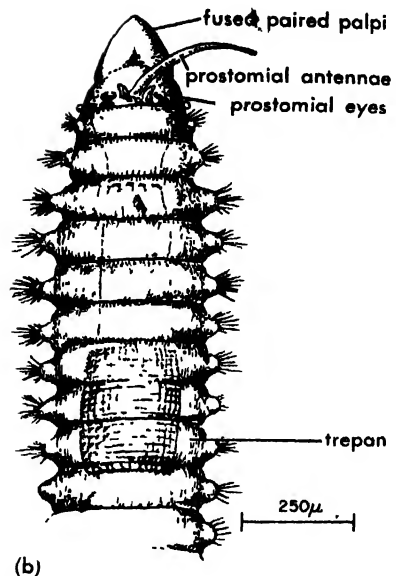
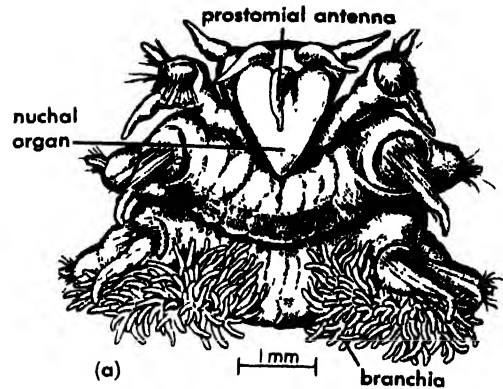


Fig. 2. (a) *Amphinome*, anterior end in dorsal view, showing the prostomium, caruncle, and first two segments. (b) *Exogone*, anterior end in dorsal view, showing the anterior end of the alimentary tract with trepan.

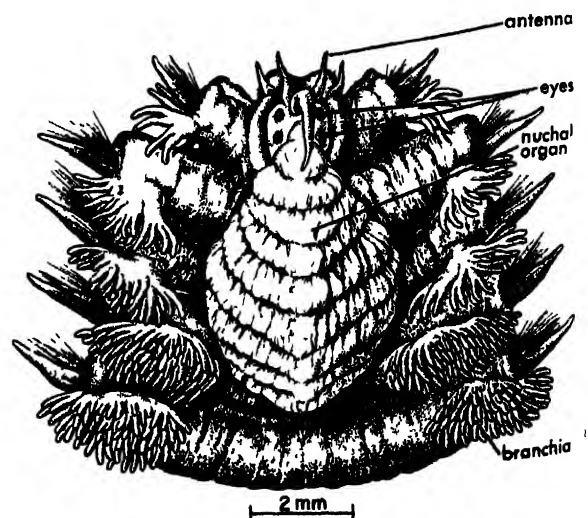


Fig. 3. *Hermodice*, anterior end in dorsal view, showing the prostomium with four antennae, large median caruncle, and parapodial branchiae.

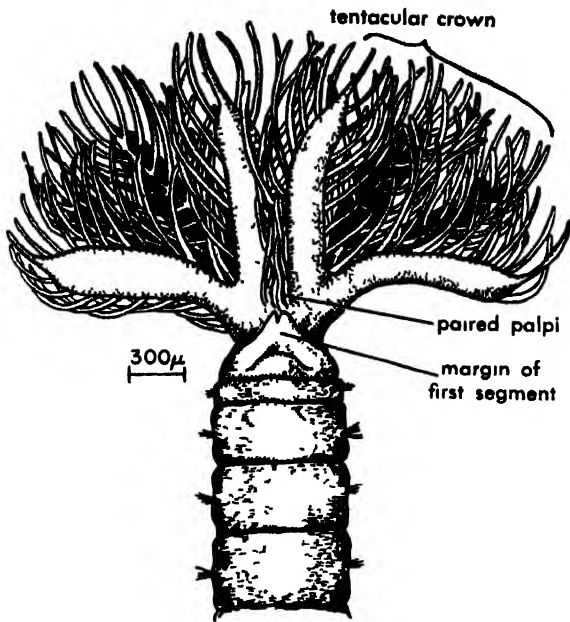


Fig. 4 *Fabricia*, anterior end in ventral view, showing feathery crown and paired palpi

operculum (Fig. 5) of a serpulid, *Chitinopoma*. It may fuse with some of the more posterior segments to form a conspicuous food-gathering organ. A pair of smaller or larger palpi (Fig. 4) are typically peristomial organs and function in food selection.

**Trunk.** The trunk in many Aphroditea consists of a definite number of segments whereas in Eunicia the number is indefinite with as few as 6, or as many as 770 segments. They may be similar to one another to form a long wormlike body or be modified into specialized regions such as thorax, abdomen, or tail. Most or all body segments have paired series of fleshy lateral expansions of the body wall, called parapodia. They may be single (uniramous) or double (biramous), consisting of a dorsal branch (notopodium) and a ventral branch (neuropodium). Each has characteristic lobes or other soft fleshy processes named according to form or function, as tentacles or cirri, elytra or scales, fringe, brachia (Fig. 3), or folds (Fig. 1).

**Parapodia.** Parapodia may be penetrated by secreted rods called acicula (entirely embedded) or by setae (emergent and retractile) of characteristic form. Setae are secretions of specialized cells within the bases of parapodia. They continue to develop throughout the life of the individual and are replaced as required. Often they form multiple series, and in many Sedentaria they may occur in thick palisaded series of many hundreds (*Serpula*, *Owenia*). Setae show extraordinary development, so that it is often possible to identify a polychaete specifically from a single characteristic type. Examples are the combs or pectinae of a nephtyid, *Aglaophamus* (Fig. 6a), and an orbiiniid, *Naineris* (Fig. 6b). *Disoma* has brushes of unique form. Harpoonlike setae or spears are conspicuous

in *Phylo*. The operculum of *Sabellaria* is strengthened with paleae which differ in outer (Fig. 6c) and inner (Fig. 6d) rows. Many Sedentaria have avicular hooks or uncini peculiar to family or generic groups such as those of terebellids (Fig. 7a), sabellids (Fig. 7b), and serpulids (Fig. 7c).

**Setae.** Setae may be simple (Fig. 6g) or composite (Fig. 6f). They may be distally hooded (Fig. 6h) or transversely ridged (Fig. 6i). The various kinds of setae function in progression for maintaining grip, for swimming, as organs of defense or offense, and in reproductive phenomena. In some pelagic polychaetes setae are lacking and parapodia are modified as paddles or oars. In some scale worms the dorsal setae are modified to form a dorsal felt, and in some polvodontids a complex glandular organ secretes the threadlike mesh which forms thickly matted tubes.

**Alimentary tract.** The alimentary tract shows many departures from a simple cylindrical tube. An anterior end or proboscis can be eversible as a soft simple or divided pouch, or as a long, richly

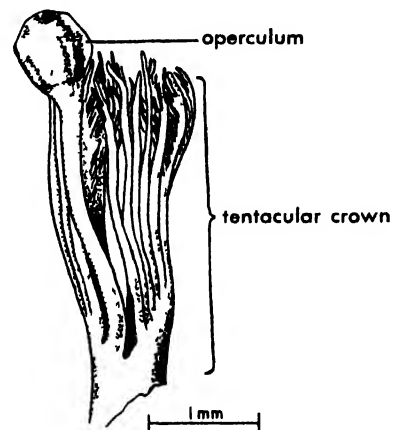


Fig. 5. *Chitinopoma*, showing tentacular crown and operculum

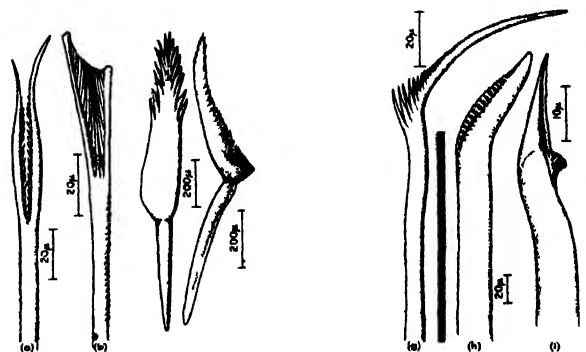


Fig. 6. Various types of setae. (a) Pectina of a nephtyid, *Aglaophamus*. (b) Pectina of an orbiiniid, *Naineris*. (c) Palea of *Sabellaria* from an outer row. (d) Inner row palea of *Sabellaria*. (e) Brushlike plumed seta from *Disoma*. (f) Composite seta of the syllid, *Exegone*. (g) Simple seta of the sabellid, *Fabricia*. (h) A hooded seta. (i) A ridged seta.

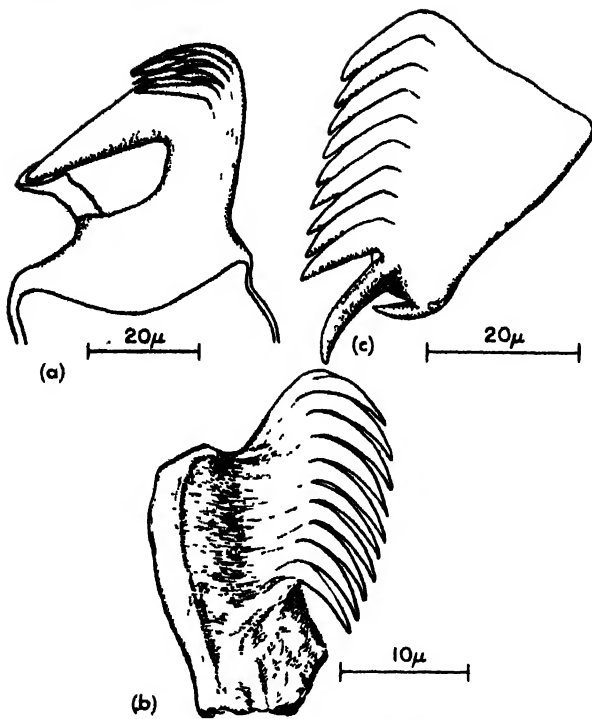


Fig. 7. Uncini of (a) *Amphitrite*, a terebellid; (b) *Fabricia*, a sabellid; (c) *Chitinopoma*, a serpulid.

adorned tube terminating distally in hard, chitinized jaws. A part of the digestive tube may be in the form of a trepan (Fig. 2b). In the Errantia it is usually more highly evolved than in the Sedentaria; in the former it is often eversible, in the latter seldom or never.

**Nervous system.** The nervous system consists of a brain or dorsal cephalic ganglion, connected to single or multiple ventral cords by circumesophageal connectives. Various ramifications of this system innervate the fleshy membranes and form networks of great complexity, especially in heteronereid epitoke stages. In many Sedentaria the peripheral nervous system is accompanied by a giant fiber system, functioning in phenomena of abrupt response.

**Respiratory system.** The respiratory system is variously developed and externally manifest as epithelial outgrowths called branchiae or gills. They are simple filaments or much divided as in *Hermodice* (Fig. 3). Typically each lobe is penetrated by vascular loops, making connections with transverse vessels. They effect a more rapid exchange of oxygen with respiratory wastes.

**Circulatory system.** The circulatory system consists of median and longitudinal dorsal and ventral vessels, continuous at either end, and with segmentally arranged transverse connectives, extending to all parts of the body. The general direction of flow is forward in the dorsal and backward in the ventral vessel. Pulsation emanates from a part of the dorsal vessel which may be modified as a heart. A cardiac or heart body, present in many Sedentaria, is muscularized, functioning to stimu-

late more rapid flow of the blood. Reduction of the vascular system to an open one occurs in many Errantia.

The blood may be colorless, reddish, greenish, bluish, or yellow. Recent studies by H. Munro Fox have demonstrated the presence of a dichroic (red-green) respiratory pigment called chlorocruorin which appears red in concentrations and green in dilutions. This is present, with red hemoglobin, in at least some species of *Serpula*. Both pigments are allied to hemoglobin of vertebrate blood and differ in their reactions, especially the degrees of affinity for oxygen and carbon dioxide. See RESPIRATORY PIGMENTS.

**Excretory system.** Excretion, or elimination of liquid wastes is accomplished by segmental organs or nephridia. In many Errantia they are closed at their inner ends and terminate in clusters of small cells called solenocytes. Most Sedentaria and some Errantia have more highly evolved nephridia, in which the inner ending is a ciliated funnel and the tube is spiralled or twisted. At maturity such nephridia may function to release gonadal products.

**Reproductive system.** The reproductive system may be simple or complex. In the simplest form the sexes are separate and the generative cells are proliferated from the coelomic walls, then shed into the coelom where the cells mature. The external form of the body often changes, either in color or surface texture, or a radical change may occur such as long swimming setae replacing normal ones. Parapodial lobes may become greatly enlarged, and many changes, both outside and inside the body, cause a metamorphosis. Familiar examples are the heteronereids of the family Nereidae, and the dimorphic phases of the Syllidae, illustrated by the polybostrichus (Fig. 8) of the male, and the sacconereis of the female. In these cases the adult is pelagic and called an epitoke, whereas its nonpelagic phase is an atoke.

In some polychaetes such as *Ophryotrocha*, a shorter and younger male phase is followed by a longer and older female, or changeover from female to male phases may result from ablation of a posterior part. In hermaphrodites both male and female elements are concurrently present, usually in different parts of the body, and cross-fertilization is sometimes necessary to ensure viable offspring.

Release of ova may be through special gonoducts or through modified nephridia, or by disruption of the body wall, as in the palolo worm. The fertilized ovum gives rise to a small spherical, ciliated larva called a trochophore, which may maintain a short or long existence in the plankton. Often the earliest stages are not planktonic, but modified to develop either in the tube of the adult or in special cocoons or egg capsules. Larval life may be short or long, varying from a few hours to months. Metamorphosis from larval to adult stage may be gradual, or abrupt as in *Sabellaria* where the anterior region is abruptly telescoped into the first several segments

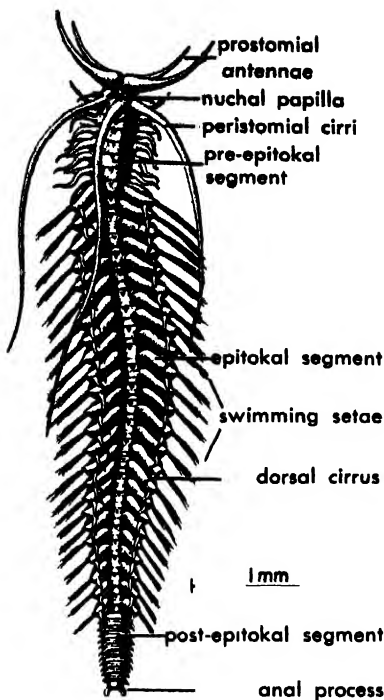


Fig 8 *Autolytus*, male polybostrichus epitoke, in dorsal view, showing modified regions of the body

to form an operculum. In some cases the larva is free and the adult is adapted to commensalism exemplified in *Spinther* (Fig. 1) in which the adult is greatly modified to live on certain kinds of sponges.

An effective but less common method of reproduction is by posterior proliferation, variously called budding, chain formation, transverse fission, and fragmentation. Numerous kinds of syllids and some serpulids are well-known examples. See REPRODUCTION, ANIMAL.

**Regeneration.** Regeneration or replacement of lost parts is very strongly developed. Polychaetes generally replace lost organs, such as parapodia, tentacular crowns, or tails, and some cirratulids can regenerate entire individuals from single segments. Autotomy, or the throwing off of damaged parts, frequently precedes regeneration. See AUTOTOMY; REGENERATION (BIOLOGY).

#### FOOD AND ECOLOGY

**Food.** Many polychaetes are deposit feeders, subsisting on nutrient particles contained in the substrata they occupy. Large amounts are engulfed by a scooping action with the aid of an eversible part of the alimentary tract. Nondigestible fractions are either expelled orally or extruded through the anal aperture. The coiled castings of *Arenicola* on some intertidal beaches are striking examples of the large amounts of sediments stirred up. Many Sedentaria are filter feeders and have highly modified food-gathering devices, such as the tentacular crowns of *Serpulea* and *Terebellidae*, and the ciliated grooves of spioniform annelids. A ciliary mechanism functions to propel food particles to-

ward the mouth. Many polychaetes, especially those with an eversible proboscis, have special capturing or holding devices. Such are the large formidable jaws and paragnaths of many Eunicia and Nereidae; they function not only to capture living prey, but to grasp and tear off sizable pieces of algae and other forms of attached life.

Most polychaetes are free-living but some are partly or entirely dependent on another animal, either as a commensal (without apparent damage to the host species) or as a parasite (more intimately associated with a host and presumably injurious to it).

Among the scale worms some genera, *Arctonoë*, *Hesperonoe*, *Lepidasthenia* and *Halosydna*, among others, are frequently associated with hosts of selected kinds, existing either near or along some part of the host which serves as a food canal. Among the Eunicia, *Iphitime* occurs in the branchial chamber of large crabs; *Labidognathus* is found in the coelom of other polychaetes; the hesionid, *Podarke*, is usually associated with certain species of asteroids, but is sometimes free-living. All Myzostomes are parasitic on or in comatulid echinoderms (sea lilies).

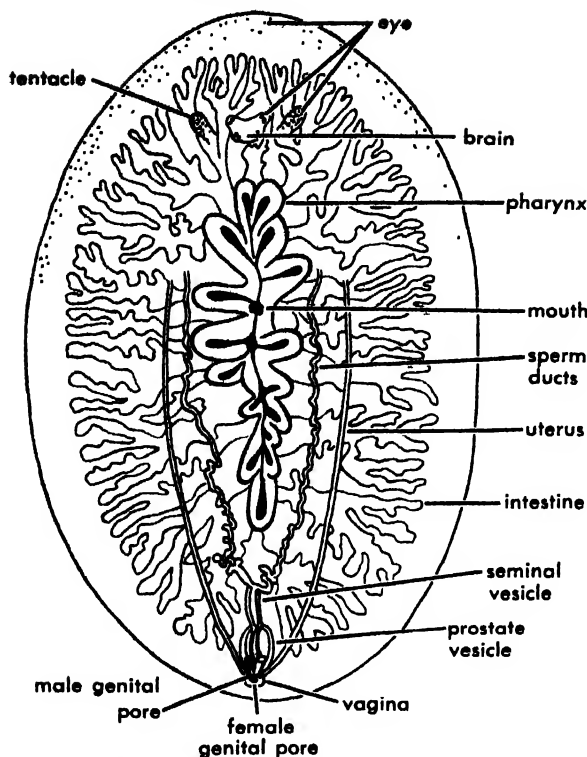
**Ecology and distribution.** Most polychaetes are marine and reach their peaks of abundance in littoral zones of all seas. They are largely stenohaline, that is, they exist in sea water of certain salt concentrations. A few, exemplified by some nereids, nephtyids, capitellids, and sabellids, are euryhaline, or can withstand the freshening effects of streams at their effluent into the sea. Fewer are fresh-water or terricolous; best known in the United States is a small sabellid, *Manayunkia leidyi*, occurring in the sediments of the Great Lakes and tributaries along the eastern Atlantic states.

Distribution of kinds is determined by such factors as latitude, depth, temperature, proximity to land masses, and kinds of sediments. The fireworms or amphinomids are largely tropical in circummediterranean areas. Eunicia and Chaetopteridae are most abundant and diverse in tropical and neotropical zones, but have many representatives of different species in colder seas. Polynoids, maldanids, nephtyids, and nereids are largely temperate. See ANNELIDA; FRESH-WATER ECOSYSTEM; MARINE ECOSYSTEM. [O.H.]

**Bibliography:** W. G. Kükenthal and T. Krumbach (eds.), *Handbuch der Zoologie*, vol. 2, pt. 2, 1931.

#### Polychaetida

A class of marine Turbellaria which are several millimeters to several centimeters in length and whose leaflike body has a central intestine with radiating branches. Most species live in the littoral zone on the bottom, on seaweed or on other objects, or as commensals in the shells of mollusks and hermit crabs. None are parasitic. Except in warm waters, they are seldom brightly colored.



*Stylochus ellipticus*. Length to 150 mm. (After A. S. Pearse, 1938)

Usually they have many eyes, and tentacles are frequently present. Frontal organs and statocysts are absent and adhesive organs are rare. The epidermis is covered with cilia and contains numerous rhabdite glands. Near the middle of the body is the mouth, followed by the plicate pharynx which opens into the central cavity of the intestine. The brain lies anterior to the pharynx, with a number of nerves radiating from it, but the two nerves which parallel the pharynx are usually the largest. Ovaries and testes are numerous and scattered, but yolk glands are lacking. Sperm ducts connect the testes with the copulatory apparatus which is variable in structure and often multiple rather than single. Insemination occurs through copulation or hypodermic impregnation. The entolecithal eggs usually accumulate in the oviducts or uteri and after fertilization pass to the exterior through the vagina. Lang's vesicle, a bursa with a long stalk, is generally present. Müller's larva, the only free larval stage known in the Turbellaria, is found in some polyclads, but is lacking in most. This larva may indicate an evolutionary link between the Turbellaria and the Ctenophora, and between the Turbellaria and the Annelida.

*Notoplana* and *Stylochus* are two of the largest and best known genera of polyclads with representatives from both coasts of North America. In particular, species of *Stylochus* are often large forms 5 centimeters or more in length. See TURBELLARIA. [E.R.J.]

**Bibliography:** R. Stummer-Traunfels, *Polycladida*, in H. G. Bronn (ed.), *Klassen und Ordnungen des Tierreichs*, vol. 4, 1930-1933.

## Polyester resins

Polymeric materials in which ester groups

O

—O—

are in the main chains. The aliphatic polyesters tend to be relatively soft, and the aromatic derivatives are usually hard and brittle or tough. The properties of either group may be modified by cross linking, crystallization, plasticizers, or fillers.


The commercial products are alkyds which are used in paints and enamels, unsaturated polyesters or unsaturated alkyds which are used extensively with fiber glass for boat hulls and panels, polyethylene terephthalate which is used in the form of fibers and films, and the aromatic polycarbonates.

This article is devoted mainly to these four products. The polydiallyl esters are frequently listed with the polyesters and will be briefly mentioned. However, their polymers are not true polyesters as defined above.

**Alkyds.** The alkyds have been in common use as coatings since World War I. In the beginning, they consisted almost entirely of the reaction products of *o*-phthalic anhydride and glycerol, and pigment. Because the functionality of the system is greater than two, a cross-linked insoluble polymer is formed. The fully cured product is quite hard and brittle. Flexible and tough materials can be produced by incorporation of monobasic acids or monohydroxy alcohols in proportions sufficient to increase flexibility but insufficient to prevent curing. Combinations of conventional vegetable drying oils and alkyd resins represent the basis of most of the oil-soluble paints. For example, by heating a mixture of dehydrated castor oil, the glycerol ester of linoleic acid, with suitable proportions of glycerol and phthalic anhydride, an oil-soluble polyester is formed. A common oil paint is produced by the addition of thinners, such as aromatic hydrocarbon solvents, a paint drier such as cobalt octoate, and pigments. By exposure to air in the presence of the paint drier, the unsaturated diene groups of the linoleic ester polymerize to yield a tough, weather-resistant coating. See DRYING OIL; POLYMERIZATION.

The drying oil-alkyd described above may be further modified by the inclusion of a vinyl monomer, such as styrene, in the original esterification process. Some of the styrene polymerizes perhaps as a graft polymer, and the remainder polymerizes and copolymerizes in the final drying or curing of the paint.

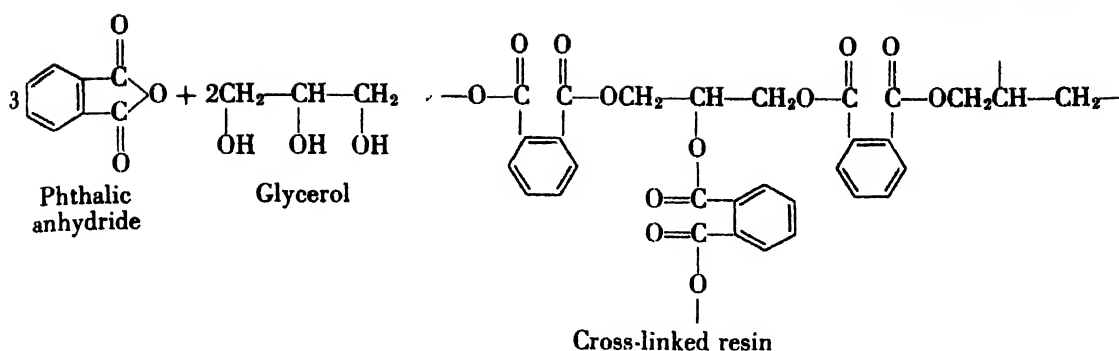
**Unsaturated polyesters.** The unsaturated polyesters were developed during and shortly after World War II. In combination with glass, they found immediate applications as panels, domes, boat hulls, and protective armor for aircraft. The compositions are distinguished by ease of fabrication and high impact resistance.



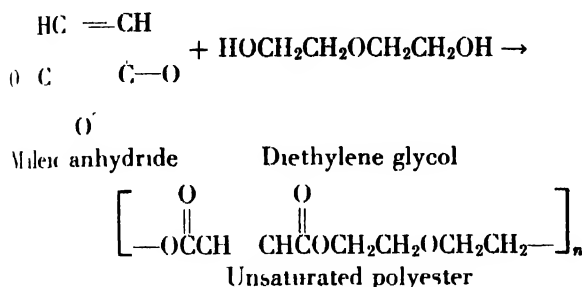
Photomicrograph of chemical reaction occurring when an acid forms a polyester compound. Photographed by Dr. Roman Vishniac using an arrangement of Leica camera equipment, high-magnification microscope, and light-polarizing equipment. (Allied Chemical Corp.)





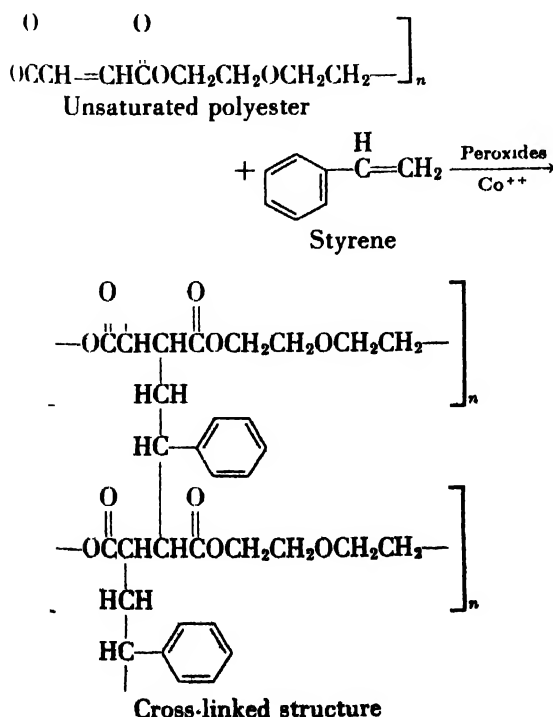


A low-molecular-weight, unsaturated polyester intermediate is first produced. The reaction of maleic anhydride with diethylene glycol is typical.



The product is a viscous oil of molecular weight 12000-4000.

The low-molecular-weight unsaturated polyester will cross link in the presence of a peroxide by copolymerization with styrene or other vinyl monomers. The unsaturated maleic group copolymerizes in essentially a 1:1 ratio with styrene. Therefore, each styrene molecule which reacts effectively joins two ester chains together to yield an insoluble cross-linking structure, such as

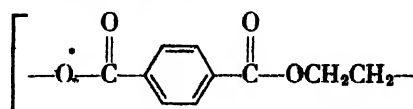


The commercially available intermediate unsaturated polyesters usually contain about 30% styrene or other vinyl monomer. On addition of a peroxide or other free-radical catalyst and a paint drier, the copolymerization starts. In this stage, the resin may be handled as a viscous fluid for a few minutes to a few hours, depending upon the activity of the catalyst. The viscous liquid may be applied to glass fiber (with a special surface treatment) in the form of matt, tow, roving, or cloth, with precautions to eliminate air bubbles and to avoid bubbles that may be caused by overheating as a result of too rapid curing. The surface of the glass fiber must have been given a special finishing treatment in advance for the polyester to adhere strongly. Glass fibers treated with a vinyl silicone or an organochrome complex are commercially available.

In the absence of the paint drier, oxygen of the air has an inhibiting effect on the curing process with the result that the surface of the product remains soft after the inner portions have hardened. In the presence of a paint drier, such as cobalt naphthenate, this skinning effect is eliminated.

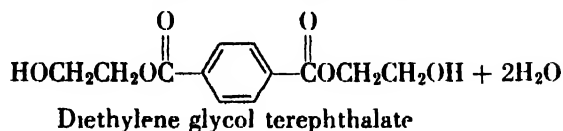
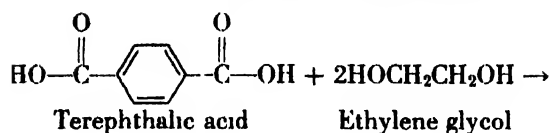
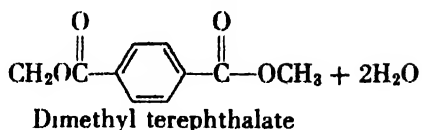
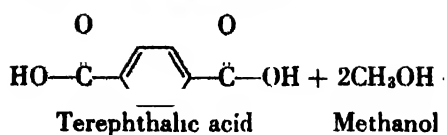
In recent years, a number of modifications of the composition described above have been made. Other acids, other glycols, and various combinations may be used to vary properties, such as flexibility, of the final product. The chlorinated derivatives have higher resistance to burning. By varying the free-radical initiator, the optimum temperature required for curing may be varied. There are thermosetting molding compositions which have glass fiber as a filler, and a catalyst which is relatively inactive at ordinary temperatures. The mixture is cross linked in the heated mold by the conventional process for thermosetting molding compounds.

**Polyethylene terephthalates.** The aromatic polyesters which have achieved general importance are the polyethylene terephthalates



which yield very strong and chemically resistant fibers and films. Polyethylene terephthalate is the principal ingredient of the polyester fibers that are available in this country and in Europe.

The preparation of the polymer involves several steps. First, the dimethyl or diethylene glycol ester of terephthalic acid is produced and isolated.



Dimethyl terephthalate is then converted to polyethylene terephthalate through ester interchange by heating with ethylene glycol in the presence of a catalyst. Further heating under vacuum of the condensate eliminates the methyl alcohol and any excess ethylene glycol and low-molecular-weight polymers, and results in the formation of high-molecular-weight, amorphous polyethylene terephthalate. If the diethylene glycol ester is utilized instead of the dimethyl ester, further heating under vacuum yields the polymer with the elimination of the excess ethylene glycol.

Ethylene glycol is obtained by the oxidation of ethylene and terephthalic acid by the oxidation of *p*-dialkyl benzenes such as *p*-xylene or *p*-cymene.

As first produced, the polymer is usually amorphous, but it readily crystallizes on reheating or on extension of the spun filaments or cast or extruded sheets. Polyethylene orthophthalate does not crystallize readily, nor does it yield useful fibers and films. Polyethylene terephthalate does crystallize readily, and has the very high crystalline melting point of 249°C. See POLYMER PROPERTIES.

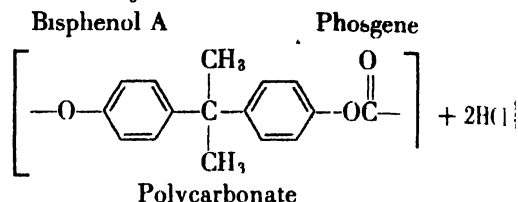
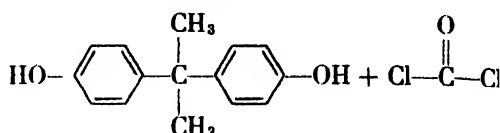
The fiber is resistant to mildew and moths. It is used frequently in combination with cotton for women's wear and men's shirts. Its chemical and heat resistance have placed it in demand for sails and cordage.

The film is tough, strong, and insensitive to moisture. It is used for special packaging, as photographic film, in electrical transformers and capacitors, and in high-strength laminates.

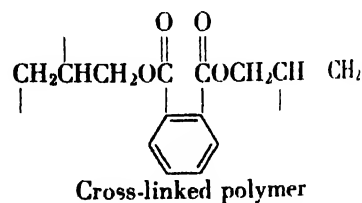
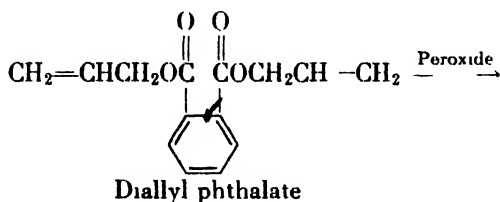
**Aromatic polycarbonates.** These are a strong, tough group of thermoplastic polymers formed most frequently from bisphenol A and phosgene. The products are distinguished by high softening temperatures, usually greater than 140°C, and high impact resistance.

The reaction between bisphenol A and phosgene leads to the polycarbonate and the evolution of hydrogen chloride. Bisphenol A is obtained by the condensation of phenol and acetone, and phosgene is produced by the reaction of carbon monoxide with chlorine.

The polymer has recently become commercially available in this country as a molding compound. It is being recommended for electrical housings, and as a replacement for metals in certain applications, such as die castings and brass and zinc bearings and bushings. Because it combines high impact strength and high softening temperature (for a thermoplastic), the product may be expected to grow into many applications.



**Polydiallyl esters.** These are polymers of diallyl esters, such as diallyl phthalate, diallyl carbonate, diallyl phenyl phosphonate, and diallyl succinate in which cross-linked products are produced by polymerization of the allyl groups, as in the case of diallyl phthalate.



Thermosetting molding compounds may be produced by careful limitation of the initial polymerization to yield a product which is fusible. Then the polymerization and curing are completed in the final molding operation.

Copolymers of diallyl phenyl phosphonate with methyl methacrylate may have a refractive index equal to that of glass. Glass fiber laminates of the product are almost clear, and are resistant to burning. See PLASTICS FABRICATION.

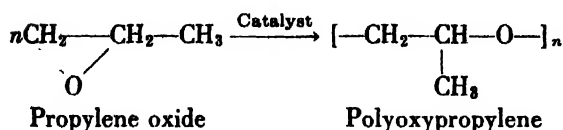
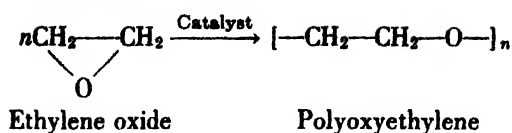
[J.A.M.; L.M.H.]

## Polyether resins

Thermoplastic materials which contain ether-oxygen linkages,  $-\text{C}-\text{O}-\text{C}-$ , in the polymer chain. Depending upon the nature of the reactants and reaction conditions, polyethers with a wide range of properties may be prepared.

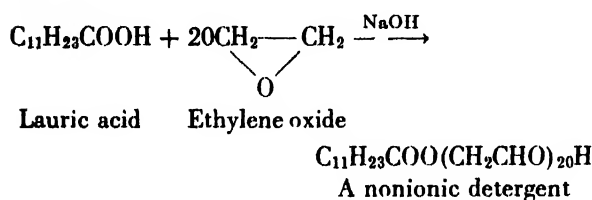
Three main groups of polyethers in use are (1) epoxy resins, prepared by the polymerization



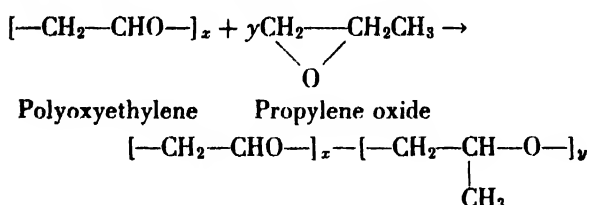


Low-to-moderate molecular-weight polyoxyethylenes vary in form from oils to waxlike solids. They are relatively nonvolatile, are soluble in a variety of solvents, and have found many uses as thickening agents, plasticizers, lubricants for textile fibers, and components of various sizing, coating, and cosmetic preparations. The polyoxypropylenes of similar molecular weight have somewhat similar properties, but tend to be more oil-soluble (hydrophobic) and less water-soluble (hydrophilic).

Nonionic surface-active agents can be prepared from  $\text{C}_{10}$ – $\text{C}_{20}$  fatty alcohols and acids by the condensation of some 5–40 ethylene oxide groups, for example,



An interesting commercial example of a block copolymer has been produced by polymerization of propylene oxide onto polyoxyethylene to yield a linear chain with sequences of the two compounds,



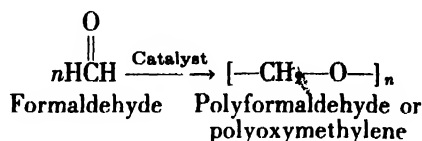
in which  $x$  and  $y$  may be in the range 10–100. The hydrophobic-hydrophilic balance necessary for surface activity is achieved because the polyoxyethylene sequence is relatively hydrophilic and the polyoxypropylene sequence is relatively hydrophobic. See POLYMERIZATION.

Oil-water emulsions prepared by use of the product have remarkable stability to both hydrophilic and hydrophobic precipitating agents.

Solid, high-molecular-weight, crystalline polymers of ethylene oxide and propylene oxide have been prepared by use of special catalysts. The polymer prepared from ethylene oxide in the presence of strontium carbonate is highly crystalline and melts at about  $66^\circ\text{C}$ . It is water-soluble, resistant to oils and greases, and is recommended for thickening and sizing applications and for extruded or cast films.

Crystalline, high-molecular-weight polyoxypropylenes with melting points up to about  $74^\circ\text{C}$  have been prepared with three groups of catalysts: (1) solid potassium, (2) complexes of ferric or stannic chloride with propylene oxide, and (3) certain metallic alkyls such as aluminum triethyl. By starting with optically active propylene oxide, an optically active polymer is produced.

**Polyoxymethylene.** Polyoxymethylene has a high molecular weight and is a very tough and strong thermoplastic material. The product has recently become commercially available, and has



promise for diverse uses in molded and extruded articles because of its high strength and toughness, and its chemical and electrical properties. It is recommended for carburetor parts, oil-resistant electrical cable sheathing, pump impellers, and water-sprinkler gears. The tendency of polyoxymethylene to depolymerize on heating has been eliminated in the commercial product, presumably by use of a polymerization system that yields inert end groups. See EPOXIDATION; PLASTICS FABRICATION; POLYMER PROPERTIES. [J.A.M.; L.M.H.]

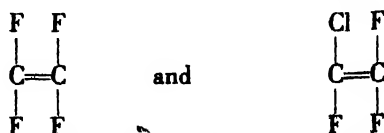
## Polyethylene glycol

Water-soluble, oily liquids, and waxes, of general formula,  $\text{HO}(\text{CH}_2\text{CH}_2\text{O})_n\text{H}$ . Their properties depend upon their molecular weights. Diethylene and triethylene glycols are used as dehydrating agents for natural gas, as textile lubricants, as humectants for glues and cork, and as starting materials for the manufacture of plasticizers and explosives. The higher polyethylene glycols (up to a molecular weight of 800) are relatively nonvolatile liquids, and find applications as heat transfer agents, lotion ingredients, and in the synthesis of nonionic surfactants. Solid polyethylene glycols are white crystalline products useful in pharmaceutical ointments, cosmetic creams, and rubber lubricants.

Polyethylene glycols with molecular weights up to 20,000 are manufactured by an alkaline-catalyzed reaction of ethylene glycol with varying amounts of ethylene oxide. See GLYCOL; POLYHYDROXY ALCOHOL. [J.T.A.]

## Polyfluoroolefin resin

A resin distinguished by resistance to heat and chemicals and by the ability to crystallize to a high degree. Two main products are the polymers of

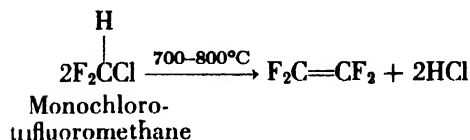
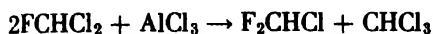
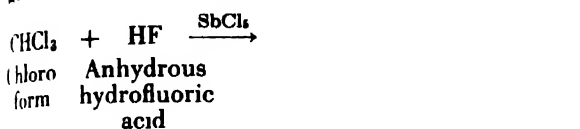


Tetrafluoroethylene      Monochlorotrifluoroethylene

Copolymers of these monomers with various fluoro-

olefins, including vinylidene fluoride,  $\text{CH}_2=\text{CF}_2$ , and hexafluoropropylene,  $\text{CF}_2=\text{CFCF}_3$ , are available. For a description of polyvinyl fluoride, see POLYVINYL RESINS.

**Polytetrafluoroethylene.** Tetrafluoroethylene can be obtained by the pyrolysis of monochlorodifluoromethane, which, in turn, is obtained from a rather complex reaction between anhydrous hydrogen fluoride and chloroform.



Although polymerization in bulk can proceed with violence, the monomer can be polymerized readily and conveniently in emulsion under pressure, using free radical catalysts such as peroxides or persulfates. The polymer is insoluble, resistant to heat and chemical attack, and in addition, has the lowest coefficient of friction of any solid. Because of its resistance to heat, the fabrication of polytetrafluoroethylene requires modification of conventional methods. After molding the powdered polymer using a cold press, the moldings are sintered at 360–400°C by procedures similar to those used in powder metallurgy. The sintered product can be machined or punched. Extrusion is possible if the powder is compounded with a lubricating material. Aqueous suspensions of the polymer can also be used for coating various articles. However, special surface treatments are required to ensure adhesion because polytetrafluoroethylene does not adhere well to anything.

Polytetrafluoroethylene is useful for applications under extreme conditions of heat and chemical activity. Polytetrafluoroethylene valve seats, packings, gaskets, and tubing can withstand relatively severe conditions. Because of its excellent electrical properties, polytetrafluoroethylene is useful when a dielectric material is required for service at a high temperature. The nonadhesive quality can sometimes be turned to advantage in the use of polytetrafluoroethylene to coat rolls to which materials might otherwise adhere.

**Polymonochlorotrifluoroethylene.** The monomer is prepared by the dechlorination of 1,1,2-trichloro-2,2,2-trifluoroethane:



Polymerization can be carried out in aqueous suspension by the free-radical process in which a combination of a persulfate and bisulfite is used as the initiator.

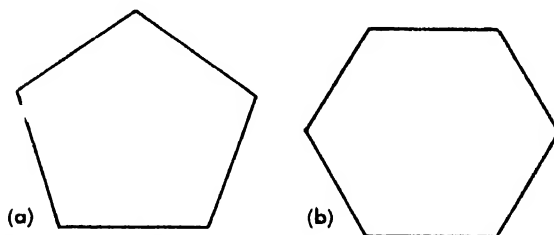
The properties of polymonochlorotrifluoroethylene are generally similar to those of polytetrafluoroethylene; however, the presence of the chlorine atoms in the former causes the polymer to be a little less resistant to heat and to chemicals. The polymonochlorotrifluoroethylene can be shaped by use of conventional molding and extrusion equipment, and it is obtained in a transparent, noncrystalline condition by quenching. Dispersions of the polymer in organic media may be used for coating.

The applications of polychlorotrifluoroethylene are in general similar to those for polytetrafluoroethylene. Because of its stability and inertness, the polymer is useful in the manufacture of gaskets, linings, and valve seats that must withstand hot and corrosive conditions. See HALOGENATED HYDROCARBON; PLASTICS FABRICATION; POLYMERIZATION.

[J.A.M.; L.M.H.]

## Polygon

A polygon of  $n$  sides ( $n > 2$ ) is a figure formed by joining an ordered set of  $n$  points called vertices by line segments called sides that connect each point to its immediate successor (where the first point is considered the successor of the last). Euclid considered a polygon to be this figure and also the plane region bounded by these vertices and sides. Modern usage considers the polygon to be any set of line segments and vertices forming a closed broken line. A polygon is simple if each side



Regular polygons. (a) Pentagon. (b) Hexagon.

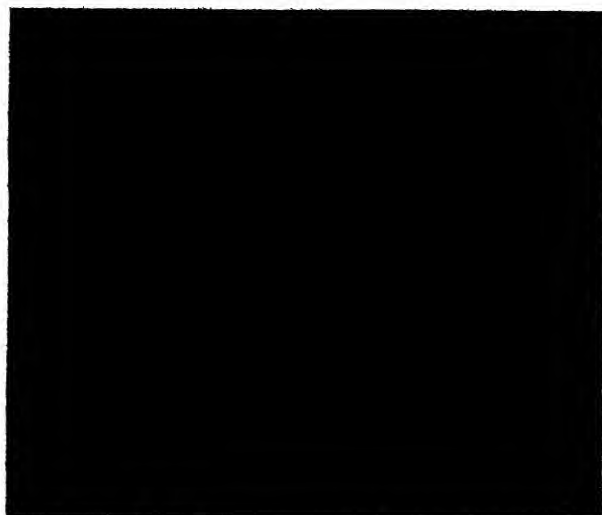
intersects only its two neighbors. It is plane or skew according as its sides do or do not lie in a plane. A simple plane polygon is convex if the polygonal region that it bounds lies wholly on one side of the (extended) line of any side. A polygon of 3, 4, 5, 6, 7, 8, 10, or 12 sides is called a triangle (3), quadrilateral (4), pentagon (5), hexagon (6), heptagon (7), octagon (8), decagon (10), or dodecagon (12). The sum of the exterior angles of any convex polygon is 360°. The sum of the interior angles is  $(n - 2) 180^\circ$ . A convex polygon is called regular if it has equal angles and equal sides. See HEXAGON; OCTAGON; PENTAGON; POLYTOPES, REGULAR; QUADRILATERAL; SQUARE; TRAPEZOID; TRIANGLE.

[J.S.F.]

## Polygonal ground

A dominant form of patterned ground characterizing fragment-bearing, gravelly, and silty soils in treeless polar and subpolar regions. Although particularly typical of regions of Arctic permafrost,





Aerial oblique view over soil polygons, north of Fairbanks, Alaska. (MATS, U.S. Air Force)

polygonal ground is also found in some high alpine areas of the middle latitudes. Related ground patterns in these regions are stone circles, nets, steps, and stripes. In each, sorted (well-defined) and non-sorted varieties are found. The polygonal forms are the most striking and always occur in large groups, never singly. The most severe climate and greatest availability of water produce the most abundant, largest, and best-sorted forms. The smallest and least-sorted types are found in less severe and more arid frost climates. Under present climatic conditions polygonally patterned ground develops in areas where the mean ambient surface temperature is less than 10°C. Where large relict forms exist outside of permafrost regions, similar paleoclimatological conditions are implied which may have bearing on Pleistocene history. Such "fossil" forms in middle latitudes suggest limits of frozen ground and permafrost associated with the climatological minima of the Ice Age.

The well-sorted type of polygonal ground is the most widespread. The sorted material grades from silt and sand to gravel and angular rock fragments. Sorted polygons have been variously termed stone polygons, stone rings, Polygonenboden Typus I, Steinnetz, and others. In these, the mesh is of dominantly polygonal form with the sorted appearance resulting from a border of stones outside of the fine material. These features range in size from a few centimeters in diameter to 10 m or more across. Unsorted forms are still polygonal, but with an absence of stone borders. Special names applied to this type are fissure polygons, mud polygons, contractional polygons, Polygonboden, Polygonenboden Typus II, Zellenboden, and other designations.

The origin of polygonal ground is not clear. Whatever the process, the patterns are created by the systematic segregation of coarse particles from the finer varieties in the surficial mantle. A clue to the process involved comes from the fact

that the largest polygonal forms are found in frost-affected areas, mainly in the polar regions. One of the chief causes seems to be local differential frost heaving. Other probable causes of sorting are desiccation as a result of aridity, and contraction resulting from extremely low temperatures. A further cause appears to be cryostatic movement through freezing-induced hydrostatic pressure which facilitates the transfer of fines to points of easiest relief. In the most intense developments, it appears that a combination of these separate processes is responsible for the striking surface patterns which such ground displays. See PERMAFROST. [M.M.M.]

**Bibliography:** A. L. Washburn, Classification of patterned ground and review of suggested origins, *Bull. Geol. Soc. Am.*, 67(7):823-865, 1956. C. Troll, *Structure Soils, Solifluction, and Frost Climates of the Earth*, U.S. Army, Snow, Ice and Permafrost Research Establishment, Corps of Engineers Translation 43, 1958.

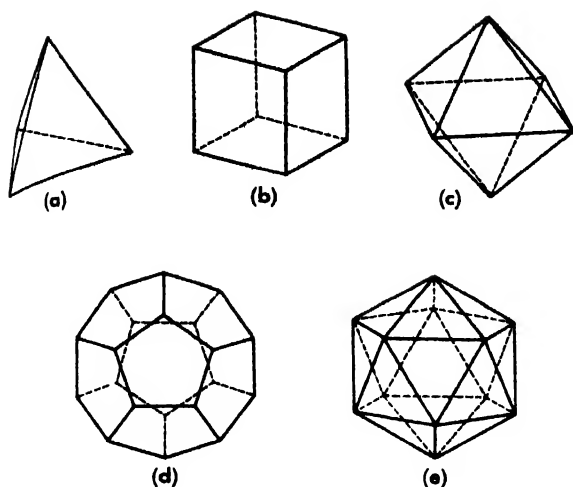
## Polygonales

An order of the plant subclass Dicotyledoneae having one family (Polygonaceae) with 32 genera and 800 species, mostly of the north temperate zone. Usually the family is characterized by swollen nodes and a stipular growth (ocrea) sheathing the stem. The fruit is a triangular or lenticular achene. In the order are many weedy plants and only a few species of economic importance. Rhubarb or pie plant (*Rheum rhaponticum*) is cultivated for its edible petioles. *R. officinale* is the medicinal rhubarb. Prince's feather (*Polygonum orientale*) and coralvine (*Antigonum leptopus*) are grown as ornamentals. The sea grape (*Coccoloba uvifera*) of tropical beaches bears edible fruit. Buckwheat (*Fagopyrum sagittatum*) is cultivated as a food plant and is prized by the apiarist as a honey plant. See BUCKWHEAT; RHUBARB; see also DICOTYLEDONEAE; EMBRYOPHYTA; PLANT KINGDOM. [P.D.S.]

## Polyhedron

A solid, all of whose boundary points lie in a finite number of planes (at least four). The bounding sections formed by these planes are plane polygonal regions called faces of the polyhedron, whose sides are called edges of the polyhedron, and whose vertices are called vertices of the polyhedron. According as the number of faces is 4, 5, 6, 8, 12, or 20, a polyhedron is called a tetrahedron (4 faces), pentahedron (5), hexahedron (6), octahedron (8), dodecahedron (12), icosahedron (20). For any convex polyhedron (but not for all polyhedrons) the numbers of vertices  $V$ , edges  $E$ , and faces  $F$  are related by the equation  $V - E + F = 2$ .

Regular polyhedrons are convex polyhedrons whose faces are congruent regular polygons, forming equal dihedral angles at each edge. If  $m$  regular  $n$ -sided polygons meet at each vertex of such a regular polyhedron, then  $mV = 2E = nF$ , and the equation  $V + F = E + 2$  implies that  $1/m + 1/n = 1/2 + 1/E$ . Either  $m$  or  $n$  or both must



The five regular polyhedrons. (a) Tetrahedron. (b) Cube (c) Octahedron. (d) Dodecahedron. (e) Icosahedron

be 3, and there are only five integral solutions of this equation. To each solution corresponds one of the five regular solids, called platonic solids, as

Name	<i>m</i>	<i>n</i>	<i>V</i>	<i>E</i>	<i>F</i>
Regular tetrahedron	3	3	4	6	4
Cube (regular hexahedron)	3	4	8	12	6
Regular octahedron	4	3	6	12	8
Regular dodecahedron	3	5	20	30	12
Regular icosahedron	5	3	12	30	20

The midpoints of the faces of a regular solid are the vertices of the dual regular solid having *F* vertices and *V* faces. Rectangular coordinates for vertices of the five regular solids can be chosen as follows, where  $\tau = (\sqrt{5} + 1)/2$ :

Regular tetrahedron:

$$(1,1,1), (1,-1,-1), (-1,1,-1), (-1,-1,1)$$

Cube  $(\pm 1, \pm 1, \pm 1)$

Regular octahedron:

$$(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)$$

Regular icosahedron:

$$(\pm \tau, \pm 1, 0), (0, \pm \tau, \pm 1), (\pm 1, 0, \pm \tau)$$

Regular dodecahedron:

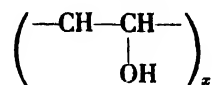
$$(\pm \tau^2, \pm 1, 0), (0, \pm \tau^2, \pm 1), (\pm 1, 0, \pm \tau^2), (\pm \tau, \pm \tau, \pm \tau)$$

See CUBE; GEOMETRY, EUCLIDEAN; OCTAHEDRON; PARALLELEPIPED; POLYTOPES, REGULAR; PRISM; PRISMATOID AND PRISMOID; PYRAMID AND FRUSTUM; SOLID (GEOMETRIC); TETRAHEDRON. [J.S.F.]

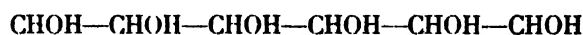
## Polyhydroxy alcohol

One of a class of compounds with more than one hydroxyl group ( $-\text{OH}$ ), each attached to separate carbon atoms of an aliphatic skeleton. This group includes glycols, glycerol, pentaerythritol, and also such products as trimethylolpropane, trimethylolpropane, 1,2,6-hexanetriol, sorbitol, inositol, and polyvinyl alcohol.

Polyols are obtained from many plant and animal sources and are synthesized by a variety of methods. 1,2-Glycols and glycerol are produced by the hydrolysis of epoxides or chlorohydrins. Formaldehyde reacts with acetaldehyde, propionaldehyde, and butyraldehyde to form, respectively, pentaerythritol, trimethylolpropane, and trimethylolpropane. Catalytic hydrogenation of sugars produces sorbitol, and 1,2,6-hexanetriol is obtained by the reduction of 2-hydroxyadipaldehyde. Saponification of polyvinyl acetate is employed in the industrial manufacture of polyvinyl alcohol



Polyols such as glycerol, pentaerythritol, trimethylolpropane, and trimethylolpropane are used in making alkyd resins for decorative and protective coatings. Glycols, glycerol, 1,2,6-hexanetriol, and sorbitol find application as humectants and plasticizers for gelatin, glue, and cork. Explosives are made by the nitration of glycols, glycerol, and pentaerythritol. The first step in the manufacture of vitamin C involves the fermentative oxidation of sorbitol to sorbose. Inositol,



is abundant in many plants and can be esterified with phosphoric acid to form inositol hexaphosphate (phytic acid), which is used in medicine as the calcium-magnesium salt. See ALCOHOL; GLYCEROL; GLYCOL; PENTAERYTHRITOL. [J.T.A.]

## Polymastigida

An order of the class Zoomastigophorea. This order, also known as Polymastigina, includes diverse genera seemingly not closely related morphologically; considerable diversity of opinion as to taxonomic status of some representatives exists. P. Grassé abandons the order. Generally they are ovoid, pyriform, or elongate, 5–350  $\mu$  long, colorless, with a thin pellicle, and are plastic. Some have cytostomes and some have undulating membranes bordered by trailing flagella. Many have an anterior rostrum behind which three (*Dallingeria*) (Fig. 1) to many (*Calonympha*) flagella emerge. Symmetry is bilateral which is sometimes obscured by a secondary radial symmetry, but one group possesses a double set of organelles, including nuclei, as mirror halves. Flagella arise from individual or fused blepharoplasts, often connected to a parabasal complex, connected with or close to an anterior nucleus, both frequently associated with an axostyle or stiffening rodlike costa.

There are at least 10 free-living fresh-water or marine genera; the remainder are parasites or commensals. Despite diverse taxonomic views, the order is of considerable ecological importance.

The termite dwelling *Trichonympha*, placed in the Hypermastigida by R. Kudo, is a bell-shaped xylophagous organism which lacks a definitive

axostyle and parabasal body. Its nucleus is centrally located and many flagella occur in longitudinal rows down the bell sides. Food is ingested posteriorly. L. Cleveland described its complicated sexual processes.

*Trichomonas buccalis*, *T. hominis* and *T. vaginalis* (Fig. 2), inhabit respectively the mouth, colon, and vagina of human beings. Some workers believe all are one species, but differ on whether they cause disease. *Tritrichomonas* (*Trichomonas*)

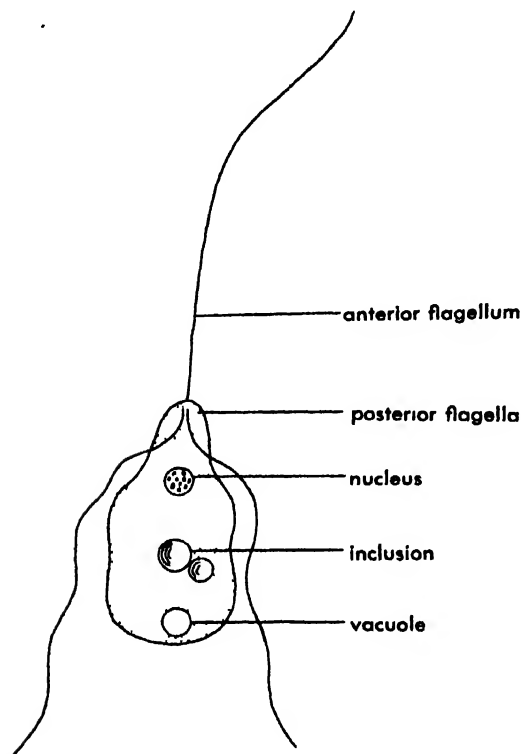


Fig. 1. *Dallingeria drysdali*.

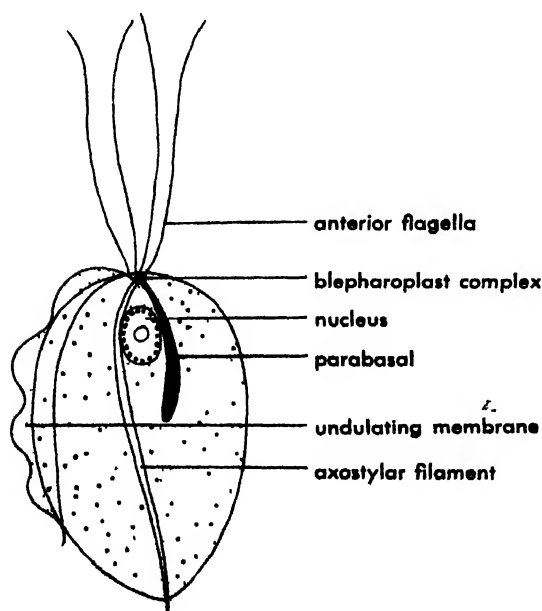


Fig. 2. *Trichomonas foetus*.

*foetus* causes abortion (Bang's disease) in cattle and *Chilomastix mesnili* is a commensal in the human colon, eating bacteria. *Giardia* (*Lambli*a) *intestinalis* occurs in the human intestinal tract. In certain cases of human diarrhea, it has been recovered with the implication that it may be the causative organism. The condition is known as giardiasis. See ZOOMASTIGOPHOREA. [J.B.L.]

## Polymer

The terms polymers, high polymers, macromolecules, and giant molecules are used to designate high-molecular-weight materials. Resin refers to an uncompounded macromolecule. Plastics, rubbers, fibers, and coatings refer to formulations of polymer with other ingredients such as fillers, pigments, plasticizers, and age stabilizers.

For a discussion of the general methods of preparation, catalytic processes, and the effects of the conditions of polymerization upon the molecular weight and molecular structure or architecture of the polymeric product, see POLYMERIZATION. For a discussion of the effect of molecular weight, molecular structure, and the conditions of fabrication upon final properties, see POLYMER PROPERTIES. For descriptions of typical synthetic polymers of the condensation type, see AMINO RESINS; CELLULOSE DERIVATIVES; PHENOL-FORMALDEHYDE RESIN; POLYAMIDE RESIN; POLYESTER RESINS; POLYETHER RESINS; POLYSULFIDE RESIN; POLYURETHANE RESINS; SILICONE RESINS; UREA-FORMALDEHYDE-TYPE RESINS. For descriptions of some important members of the addition-type synthetic polymers, see HYDROCARBON RESIN; POLYACRYLATE RESIN; POLYACRYLONITRILE RESIN; POLYFLUOROOLEFIN RESIN; POLYOLEFIN RESINS; POLYSTYRENE RESIN; POLYVINYL RESINS.

Natural products, rubbers, fibers, and paints are treated in other articles. Biological polymers, as in enzymes and living tissues, are treated in the appropriate articles. See CELLULOSE; FIBER, MAN-MADE; FIBER, NATURAL; PROTEIN; RUBBER; SURFACE COATING; TERPENE.

The first modified natural polymers, cellulose nitrate and casein-formaldehyde, were commercially produced about 1860, and the first fully synthetic polymer, phenol-formaldehyde, was made about 1910. The major development of present polymer science and technology has taken place since about 1920, while the production of polymeric materials in the United States has grown from a few million pounds to several billion pounds in the same time period.

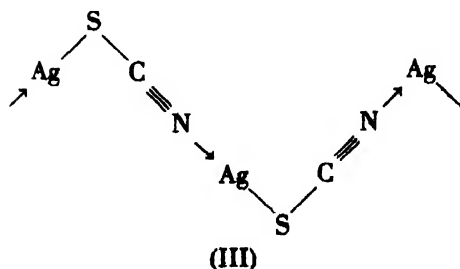
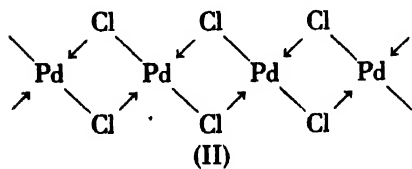
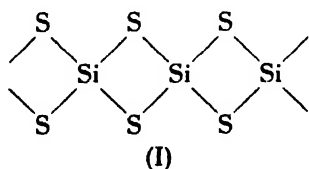
Interest in the synthesis of products similar to natural products, but possessing more useful properties, has been continually stimulated by the successful synthesis of polyamide fibers and a rubber equivalent to natural rubber, and by increasing understanding of the nature of proteins, carbohydrates, and enzymes in living tissues. The need for polymeric materials which can be easily shaped, which possess high resistance to heat and to chemicals, and which have high strength has initiated a rapidly growing interest in inorganic-

organic or completely inorganic polymers. These interests have emphasized the need for further understanding of the principles of valence, intermolecular attractive forces, and the properties of crystals and other heterogeneities dispersed in amorphous matrices. See PLASTICS FABRICATION; POLYMER, INORGANIC. [J.A.M.; L.M.H.]

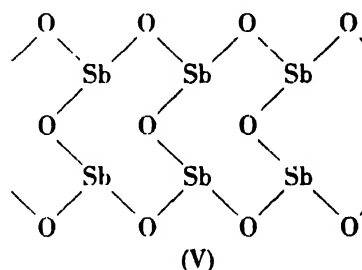
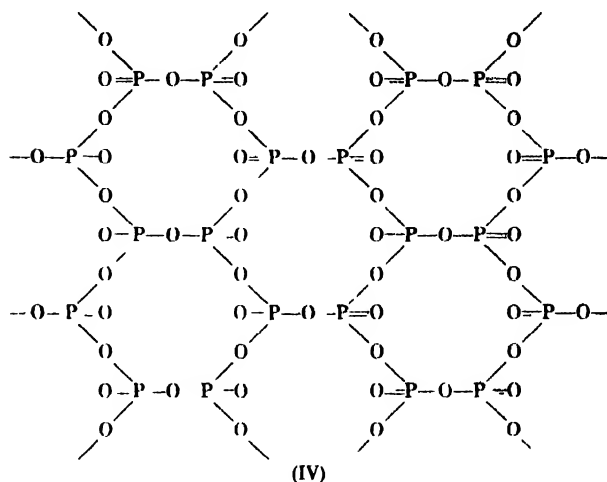
**Bibliography:** P. J. Flory, *Principles of Polymer Chemistry*, 1953; *Modern Plastics Encyclopedia*, 1960; W. J. Roff, *Fibers, Plastics, and Rubbers*, 1956; A. X. Schmidt and C. A. Marlies, *Principles of High-Polymer Theory and Practice*, 1948.

## Polymer, inorganic

A polymer whose molecular architecture is inorganic rather than organic. Interest has been directed toward synthetic inorganic polymers because of their ability to withstand high temperatures and other extreme conditions such as found in rocket technology. This property of high thermal resistance contrasts sharply with most organic polymers. Modern techniques such as x-ray analysis have shown that many simple inorganic compounds are macromolecular in nature. The silicates, in which oligometric and polymeric structures abound, are the best-known group of such inorganic compounds (see SILICATE; SILICATE MINERALS). Well-known examples of these silicates which find wide technological application are sheetlike mica and fiberlike asbestos. Particularly important are the relationships between molecular structure and physical properties, for example, the lubricating action of graphite and molybdenum disulfide. Other examples of inorganic macromolecules (as yet without technological application, however) are as follows: silicon disulfide (I), palladium chloride (II), and silver thiocyanate (III).

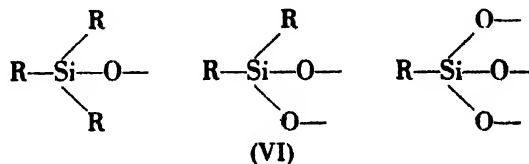


Silicon disulfide has a tetrahedral arrangement of sulfur atoms around the silicon atom, palladium chloride has a planar arrangement of chlorine atoms around the palladium atom, while silver thiocyanate consists of zigzagging chains. Other types of polymeric species are the orthorhombic phosphorus pentoxide (IV) and one modification of antimony trioxide (V):

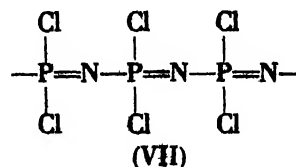


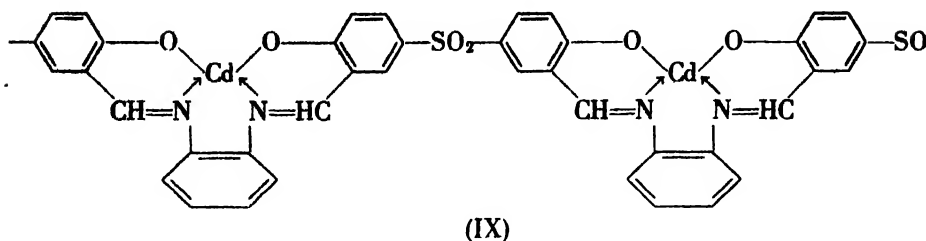
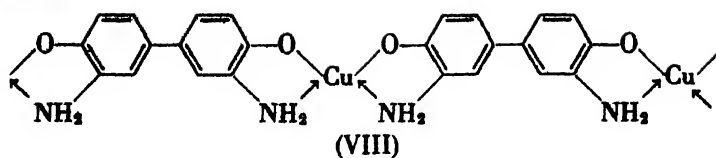
The fact that many of these types of polymers exist only in the solid state and react with solvents has directed attention toward other classes.

Perhaps best known of all the synthetic polymers based on inorganic molecular structures are the silicones, which are derived from the following basic units (VI):



These units are put together in various proportions and arrangements, and, in conjunction with different groupings R, they give a wide variety of ma-





terials varying from oils to waxes, resins, and elastomers. Silicones have found a wide variety of uses as hydraulic and dielectric fluids, lubricants, antifoaming agents, and mold-release agents, and in addition have been incorporated in many waxes and polishes. The resins can be used as electric insulators, and the silicone elastomers can be used at both higher and lower temperatures than most natural and synthetic rubbers. Other inorganic polymers which have attracted a good deal of attention are the phosphazenes (phosphonitric chlorides),  $(\text{PNX}_2)_n$ , and the phosphinoborines,  $(\text{R}_2\text{PBR}'_2)_n$ . The elastomer (VII) derived from the chlorophosphazenes has a physical resemblance to natural rubber and plastic sulfur, but it is hydrolytically unstable.

Linear silicones, chlorophosphazenes, and phosphinoborines transpose, on heating, into lower-molecular-weight cyclic forms, a tendency which has not so far been overcome. Other systems where different elements such as boron, phosphorus, aluminum, silicon, and tin are linked by means of oxygen are still in the relatively early stages of investigation.

In addition, polymeric metal coordination compounds have been made from metal ions and low-molecular-weight polydentate ligands (VIII), as well as by adding metal ions to already existing polydentate macromolecules (IX). Most of the inorganic polymers are prepared by condensation rather than by addition polymerization because of the relative lack of suitable monomers for the latter process. In this respect, the apparent scarcity of polymerizable multiple bonds in elements other than carbon is noteworthy. A few exceptions to this paucity occur, for example, in sulfur trioxide, gaseous selenium dioxide, and similar compounds.

Work on inorganic polymers has been hampered by the relative lack of knowledge of even the simplest starting materials because of the long period of neglect which inorganic chemistry has suffered in preference to organic chemistry. Difficulties in obtaining the required degree of purity in the starting materials also is a hampering factor. These factors are accentuated in many cases by the

lower volatility of the monomers, and their higher reactivity, especially toward hydrolysis and oxidation, compared with similar organic compounds.

It should be noted that while resistance to high temperatures is one of the aims in inorganic polymer research, the whole field holds promise for many other types of applications as, for example, adhesives, textile finishes, and fertilizers. See CHELATION; INORGANIC CHEMISTRY; POLYMER, POLYMER PROPERTIES; POLYMERIZATION. [R.A.SH]

**Bibliography:** R. A. Shaw, Inorganic polymers, *New Scientist*, 8(213):1603-1605, 1960; R. A. Shaw, *Inorganic Polymers*, in *High Temperature Resistance and Thermal Degradation of Polymers*, Soc. Chem. Ind. (London), Monograph 13, 1961; D. B. Sowerby and L. F. Audrieth, Inorganic polymerization reactions, *J. Chem. Educ.*, 37:2-10, 86-91, 134-137, 1960; A. F. Wells, *Structural Inorganic Chemistry*, 2d ed., 1950.

## Polymer properties

The properties of polymeric materials, which are determined by the molecular properties of the macromolecules, the type of formulation involving plasticizers or fillers, the conditions of fabrication in which molecular orientation or crystallization may be induced. Properties also depend on the temperature and elapsed time of the measurement.

**Molecular properties.** These include molecular size and weight, molecular structure or architecture, molecular-weight distribution, polarity, and flexibility of the polymeric chains (or chain segments between cross-links in cured or vulcanized polymers). Molecular properties taken together determine the attractive forces between the molecules, and the general behavior of the polymer.

**Molecular weight and distribution.** The desirable properties of high polymers (strength and resistance to solvents) increase rapidly with increasing molecular weight in the low ranges of molecular weight, and more slowly in the high ranges. On the other hand, the melt viscosity increases with molecular weight in the opposite manner, slowly at first and rapidly in the higher-molecular-weight range. The ease of fabrication (molding,

and shaping) of polymeric compositions varies inversely with the melt viscosity; that is, the materials become increasingly difficult to mold or extrude at very high values of molecular weight. The optimum molecular weight of a polymer frequently varies for different applications and different methods of fabrication. Commercial products are usually available in several molecular-weight ranges. In all cases, however, the optimum molecular weight must be selected on the basis of the best compromise for desirable properties and ease of fabrication. Techniques of measuring molecular properties and methods of fabrication are briefly discussed at the end of this article.

The molecular-weight distribution represents the particular combination of material of low, medium, and high molecular weight in a given product. The presence of low-molecular-weight portions may lower the softening temperature range of the product, make it more subject to attack by solvents and chemical agents, and lower the melt viscosity or increase the ease of fabrication. The lower-molecular-weight portions behave more or less as plasticizers. The presence of very high molecular-weight portions increases the melt viscosity but it does not have pronounced effects on the physical properties.

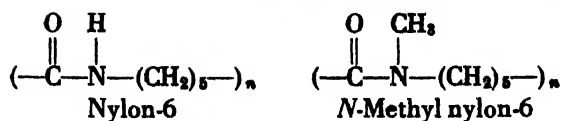
Frequently, the portions of very high molecular weight are actually insoluble, being highly branched or cross-linked. The insoluble portions show up as "fish-eyes" in films and discontinuities in fibers. They also give undesirable nerve (elastic memory) characteristics to compounded, uncured rubber.

**Attractive forces.** The atoms in the chains are held together by primary valence bonds. If it were possible to apply a force of tension only to the primary valence bonds, then a tensile strength of more than 2,000,000 psi would be observed. In reality, however, under tension the molecules slip past one another and the resistance to that slippage is due to the effective attractive forces (van der Waals forces) between the molecules. See CHEMICAL BINDING.

The effective attractive forces are the result of the polarity of the groups in, or attached to, the chains, and of the degree of fit between chains. Strongly polar groups, such as those containing oxygen, nitrogen, and sulfur, and other polar atoms exert the strongest attractive forces. Bulky side groups attached to the main chains stiffen the chains and also by their bulk may prevent a close fit between chains.

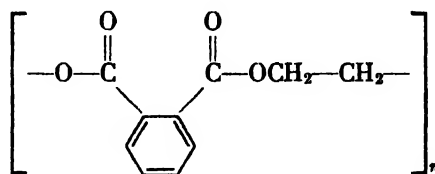
However, if the bulky side groups are arranged in a stereoregular way (as in isotactic polymers), then close fit is possible and crystallization may occur.

An example of the strong attraction of polar groups is found in the following comparison of the properties of polyamides and *N*-methyl polyamides. Nylon-6 is relatively high melting, hard, strong, and insoluble, whereas the *N*-methyl derivative is

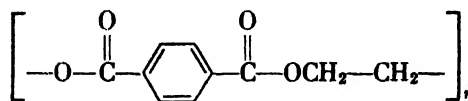


lower melting, less strong, and more readily soluble. The strength of the hydrogen bond between the >N—H group and the oxygen atom of an adjacent chain is some 10 kcal/mole. By replacing the >N—H by >N—CH<sub>3</sub>, the intermolecular attractive forces are reduced to perhaps 2–3 kcal/mole.

**Fit.** An example of the effect of the regularity of chains on the fit between them is seen in the comparison of the properties of polyethylene orthophthalate and polyethylene terephthalate.



Polyethylene orthophthalate



Polyethylene terephthalate

The para derivative, polyethylene terephthalate, has the more regular structure. It softens at higher temperatures, crystallizes more readily, and is stronger and less soluble than the other polymer. Although the chemical nature and polarity of the two polymers are identical, the effective intermolecular attractive forces between adjacent chains are greater for the para product because its structure allows a closer interchain fit.

**Crystallinity.** When a close fit between chains is possible, crystallization can take place spontaneously or by drawing or cooling. The crystals represent a configuration of minimum distance between molecules. Because the long, threadlike, coiled and entangled chains are never fully untangled, even on drawing, crystallization never reaches completion. Crystalline polymers may contain up to 90–95% crystallinity, with the crystals embedded in and exercising an effect upon the remaining amorphous polymer. The behavior of a partially crystalline polymer is somewhat like that of an amorphous polymer containing a finely dispersed filler or strengthening pigment, and for rapid stresses, it is somewhat like that of a chemically cross-linked product.

Amorphous products which are strongly oriented by cold drawing, but which are not thereby crystallized, show evidence of closer fit between chains by increased strength in the direction of stretching.

**Flexibility.** The flexibility of linear polymer chains and of the segments between cross-links in cured products is decreased by the presence of



polar groups and regularity in the molecular structure. A nonpolar, irregular chain should be the most flexible. Products containing highly flexible chains are rubbery, soft, and nonbrittle, with relatively high resistance to impact and to tear.

The limited type of Brownian motion that is possible in polymer chains and the effective attractive forces between chains are highly temperature-dependent. If the temperature is increased, the rubbery qualities increase. The second-order transition temperature or glass point is, in somewhat arbitrary terms, that temperature below which the chain segments are relatively immobile and the product has glasslike properties, being brittle and hard. The first-order transition temperature is the crystalline melting point. Above the glass point, the chain segments have relatively high Brownian motion, and the product is leathery to rubbery. Thus with variation of temperature and treatment, many polymers may exist as rubbers or gums, hard glassy solids which may be partially crystalline, or strong fibers and films which are frequently highly crystalline and have high softening points.

**Behavior under stress.** The illustration indicates the type of stress-elongation behavior shown by a typical glassy or crystalline polymer, and a typical rubber or amorphous polymer capable of strengthening or crystallization on orientation.

The nature of the properties depends also upon the elapsed time of the measurement. In high-speed vibration, rubber is stiff and acts more like a plastic glass. Under long-term tension, some glasslike plastics will simply flow as liquids.

The cross-linked polymers are usually substantially stronger than corresponding linear products; however, their strengths fall far short of the hypothetical 2,000,000 psi. Linear polymers break by overcoming the intermolecular attractive forces and possibly to some extent by the rupture of primary bonds. Cross-linked polymers break by the latter process. In both cases, the break probably starts at a fault in the sample.

**Elasticity.** There is a profound difference in the elasticity of rubbers, of hard brittle plastics, and of strong, partially crystalline fibers. When rubber is stretched, the flexible, coiling segments between

the cross-links are straightened. On release of the stretching tension, the segments resume their random coiling condition. An increase in temperature increases the freedom of the segments to seek their random condition and thereby increases the restoring force for a particular elongation. In crystalline solids, and in polymeric glasses and fibers at temperatures below their glass points, the elasticity of the Hookean type is small and is due to the bending of bonds. On increasing the temperature, the restoring force is lessened. On extension of polymeric materials, both rubberlike and Hookean elasticity may be present, and there may be some cold flow or permanent slippage of the molecules. On release of the external force, there may be an immediate recovery of the Hookean deformation and some of the rubberlike deformation, a slow recovery of some of the rubberlike deformation (in cases in which free coiling and uncoiling of the segments is impeded by polar groups), and nonrecovery of the deformation due to slippage of the molecules, or cold flow.

The term elastic memory applies to cases in which a polymer is deformed at an elevated temperature, as in shaping of a sheet into a dome, and then is cooled before the tangled chains have reached an equilibrium condition in the new shape. Strains are said to be frozen in, and at a later time, especially if the product is warmed, these strains will cause the product to assume a distorted shape.

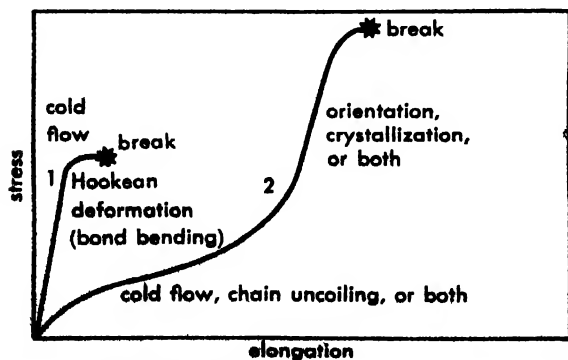
**Compounding.** Plastic masses, rubber formulations, coatings, and other polymeric compositions may contain age inhibitors, strengthening and coloring pigments, and plasticizing or softening agents. Roll mill, sigma blade, and dough mixers are generally employed to mix the resin with plasticizers and pigments at moderately elevated temperatures. 50–115°C.

**Additives.** The addition of plasticizers to a polymeric material causes it to be softer and more rubbery in character. Plasticizers are held in association with the polymer chains by secondary valence forces. They separate the molecules, thus reducing the effective intermolecular attractive forces.

Finely divided polar substances may act as strengthening fillers for plastics. The filler surfaces can adsorb links of several polymer chains and produce the effects of cross-linking. Filled polymeric compositions are frequently harder, stronger, more resistant to abrasion, and less elastic than the unfilled products. The effect of strong fiber fillers (glass, sisal, or other materials) on the tensile properties of several polymers is impressive. Products with tensile strengths of 50,000–100,000 psi can be obtained.

Age inhibitors are almost always incorporated in polymeric compositions. Oxygen, ozone, light, and electric discharge produce free radicals which cause degradation of the polymer chains. Free-radical inhibitors and light-masking agents are therefore commonly used. See INHIBITOR (CHEMICAL).

**Polyblends.** In completing a discussion of the effect of molecular-segment characteristics on the



Stress-elongation behavior of polymers. Curve 1, glassy or crystalline polymer. Curve 2, rubber or amorphous polymer.

properties of polymers, mention should be made of the interesting characteristics of polyblends. By the addition of small amounts of a rubbery polymer to a polymeric glass, the impact strength of the latter is substantially increased. The rubbery polymer is not truly compatible and exists as a finely dispersed separate phase. The phenomenon has somewhat similar counterparts in inorganic glass technology and in physical metallurgy. The study of polyblends demonstrates the growing interest in the solid-state physics and physical metallurgy of macromolecules.

**Fabrication.** Polymer formulations can be fabricated into useful forms or articles by a variety of methods.

In the use of molded thermosetting compositions in which heat and pressure are required for the production of sound articles, some form of compression molding is employed. The physically compacted composition containing the resin in the fusible stage is forced into a mold cavity of the desired shape and is held under heat and pressure until the curing or vulcanization is complete. When high pressure is not required, as in the preparation of epoxy compositions, polyester-styrene-glass fiber compounds, and in various fast-curing resin laminates, it is still desirable to use moderate pressure to obtain a uniform molding.

Various forms of injection molding are used for the shaping of thermoplastic (permanently fusible) compositions. The composition is first softened temporarily by forcing the compounded resin granules through a heating chamber, after which it is driven into a relatively cold mold. Under proper conditions, the resin remains soft long enough to fill the mold completely and then rapidly hardens as it cools. Variations of the injection molding process are used in the extrusion of films, rods, and pipe, and the spinning of fibers.

Films are also produced by extrusion of a tube into which air is forced. The process is called bubble extrusion. The pipe expands, because of the air pressure, to a wall thickness equivalent to the film thickness desired. The walls of the expanded bubble are pressed together in nip rolls and later the large, thin-walled, collapsed pipe is slit to yield flat film. Films and sheets are also produced by calendering in which the hot resin is forced between tightly fitted rolls.

In the casting process, fluid compositions are poured into molds of the desired shape and then allowed to cool or cure. This process is used for the production of foams and in encapsulation, such as in the protection of electronic components or the mounting of biological specimens. In broad terms, coating, slush-molding, and painting may be considered to be casting operations.

In the shaping of thermoplastics particularly, the conditions of molding (temperature, time, and pressure) have a marked effect on the properties of the final product. Uneven cooling produces strains and the flow of the plasticized resin can cause some orientation of the molecules. In bubble molding of film, biaxial orientation is produced. In the

drawing of films and fibers, uniaxial orientation results, although films may be partially stretched with equipment similar to the Tenter frames of the textile industry. Orientation results in a substantial increase in the strength of the product in the direction of stretching; in many polymers, crystallization is induced during cold-drawing or stretching. On occasion, discontinuities or areas of strain are produced by partially orienting the molecules, and the product becomes more subject to stress cracking in the presence of solvents or other agents. Thus the optimum condition of a product results from a judicious combination of mechanical treatment and thermal annealing. See PLASTICS FABRICATION.

**Measurement of molecular properties.** Because of the random, statistical nature of the polymerization process, a distribution of chain lengths, that is, molecular weights, is always formed, and measured molecular weights are necessarily average values. Because of the very high molecular weights of macromolecules, the common methods of measuring properties of solutions (for example, the freezing point lowering) are frequently not suitable. Among the colligative methods, the osmotic-pressure procedure is most useful and gives number-average molecular weights of reasonable accuracy up to about 500,000. The boiling-point elevation method is suitable for molecular weights up to 5,000-10,000.

The fact that the amount of light scattered by a solution is a function of the molecular weight of the dissolved particles has provided a means for the determination of weight-average molecular weights up to several million. The light-scattering measurement can also yield valuable information regarding the shape of the molecule in solution.

The determination of molecular weight and also molecular-weight distribution can be accomplished by use of the ultracentrifuge in which the rates of settling of particles in intense centrifugal fields are measured.

The intrinsic viscosity of a polymer in solution (the viscosity which the unassociated polymer molecules give to the solution) is a function of the molecular weight and is very easily measured. Intrinsic viscosity is commonly used for control purposes, and the values can be converted into molecular weight by calibration with osmotic pressure, light scattering, or sedimentation measurements.

The second-order transition temperature is a time-dependent function. Normally, it is measured by noting the temperature at which the slope of a flexibility-temperature curve changes abruptly. See COLLOID; FIBER, MAN-MADE; POLYMER; POLYMERIZATION; SCATTERING (ELECTROMAGNETIC RADIATION); ULTRACENTRIFUGE. [J.A.M.; L.M.H.]

**Bibliography:** P. D. Ritchie, *A Chemistry of Plastics and High Polymers*, 1949.

## Polymera

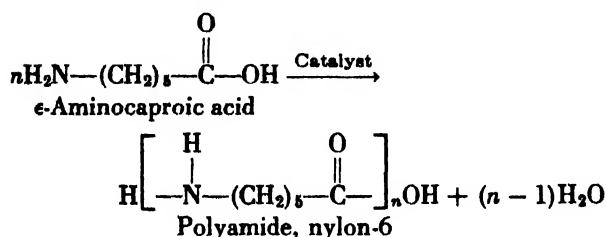
A division of the phylum Vermes proposed by O. Bütschli in 1910, with the rank of a subphylum. The Polymera are equivalent to the phylum An-

nolida. The Amers and Oligomera are the other subdivisions of the Vermes which were recognized. See AMERA; OLIGOMERA. [C.B.C.]

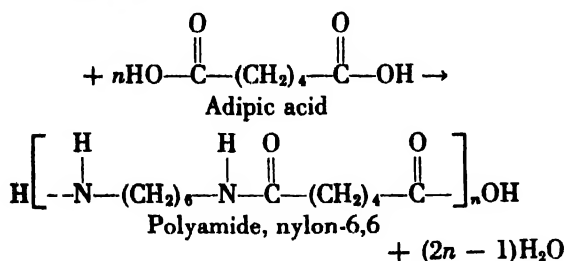
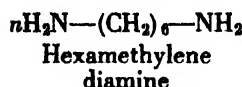
## Polymerization

The linking of small molecules (monomers) to make larger molecules. Polymerization requires that each small molecule have at least two reaction points or functional groups. There are two distinct types of polymerization processes, the condensation polymerization in which the chain growth is accompanied by elimination of small molecules such as  $H_2O$  or  $CH_3OH$ , and addition polymerization in which the polymer is formed without the loss of other materials.

An example of the condensation process is the reaction of  $\epsilon$ -aminocaproic acid in the presence of a catalyst to form the polyamide, nylon-6:



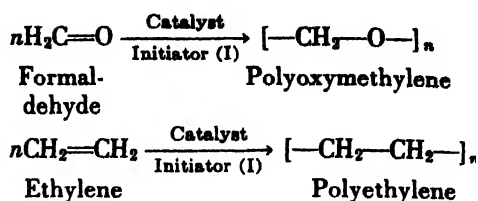
The repeating structural unit is equivalent to the starting material minus H and OH, the elements of water. A similar product would be obtained by the reaction of a diamine and a dicarboxylic acid. In both cases, the molecules formed are linear because the total functionality of the reaction system



(functional groups per molecule) is always two. However, if a trifunctional material, such as a tricarboxylic acid, were added to the nylon-6,6 polymerizing mixture, a branched polymeric structure would result, because two of the carboxylic groups would participate in one polymer chain, and the third carboxylic group would start the growth of another. At high conversion, these chains could become bridges between linear chains and the polymer would then be cross-linked.

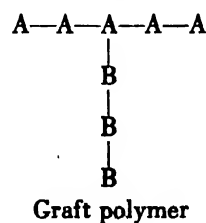
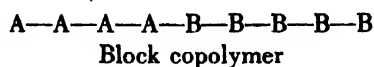
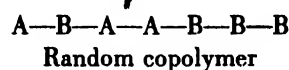


Some examples of addition polymerization are

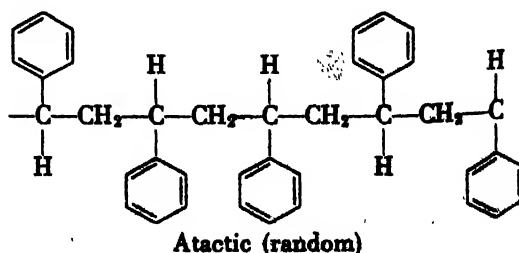


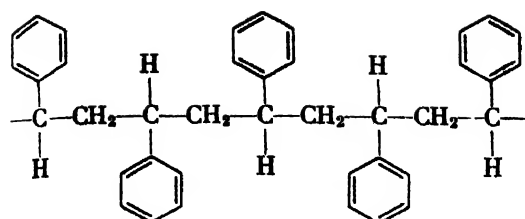
The structure of the repeating unit is the difunctional mer. In the presence of catalysts or initiators, the monomer yields a polymer by the joining together of  $n$  mer links. If  $n$  is a small number, 2-10, the products are dimers, trimers, tetramers, or oligomers, and the materials are usually gases, liquids, oils, or brittle solids. In most solid polymers,  $n$  has values ranging from a few score to several hundred thousand, and the corresponding molecular weights range from a few thousand to several million. The end groups of these two examples of addition polymers are shown to be fragments of the initiator.

If only one monomer is polymerized, the product is called a homopolymer. The polymerization of a mixture of two monomers of about equal reactivity leads to the formation of a copolymer, a polymer in which the two types of mer units have entered the chain in a random fashion. If chains of one homopolymer are chemically joined to chains of another, the product is called block or graft copolymer:

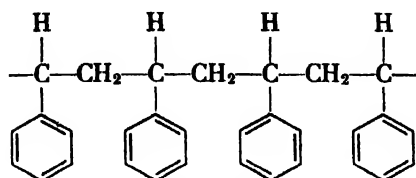


Isotactic and syndiotactic (stereoregular) polymers are formed in the presence of complex catalysts. The groups attached to the chain in a stereoregular polymer are in an ordered arrangement. The regular structures of the isotactic and syndiotactic forms make them capable of crystallization. The crystalline melting points of isotactic polymers are substantially higher than the softening points of the atactic product.





Syndiotactic (alternating)

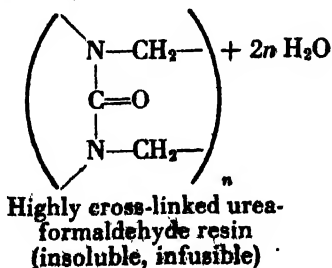
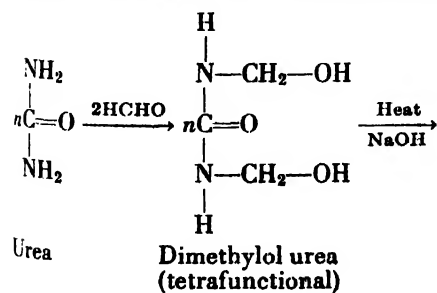


Isotactic

The production of *cis*-polyisoprene, "synthetic natural rubber," is also an example of stereoregular polymerization.

### CONDENSATION POLYMERIZATION

The formation of macromolecules by the condensation process, as in the production of polyamides, polyesters, and polysulfides, requires the elimination of small molecules as noted above; at the same time, strongly polar and strongly attracting groups are produced. In the equimolar reaction of a dicarboxylic acid with a diamine, as an example, after 50% of the groups have reacted, the average degree of polymerization (DP) of the polymer formed is 2; after 90% reaction, the DP is 10, and after 99.5% reaction, the DP is 200. If the molecular weight of the repeating unit is 100, then the average molecular weight of the polymer is 20,000. In order to obtain a high-molecular-weight product, the small molecules formed during the condensation must be removed in order for the reaction to approach 100% completion. In practice, the condensations are usually started under moderate conditions of temperature and pressure and completed at high temperature and low pressure to yield linear products in the molecular-weight range of



about 5,000–30,000. The linear products, thermoplastic and condensation resins, are used in fibers, films, coatings, molding compounds, and adhesives.

Useful condensation polymers which are highly cross linked are prepared from low-molecular-weight polyfunctional reaction systems. The condensations of formaldehyde with phenol and with urea, and the condensation of phthalic anhydride with glycerol and others are discussed in the articles on specific products. The reaction of formaldehyde with urea will serve as an example.

In the intermediate compound, dimethylol urea is tetrafunctional. The H— and HO— groups on one molecule react with the HO— and H— groups on other molecules to form water. Each urea unit finally becomes bound to other urea units via four methylene (—CH<sub>2</sub>—) bridges. Because the final product is cross linked and infusible, the final shaping operation must coincide with the final curing or cross linking. In practice, the soluble, low-molecular-weight intermediate condensate could be isolated and mixed with strengthening fillers, coloring pigments, and curing catalysts to yield a molding powder. By subjecting the urea-formaldehyde molding powder to heat and pressure in a mold, the curing reaction takes place, some of the water is driven off as steam, and some is adsorbed by the filler. The molded object, such as a radio cabinet or lamp shade, is now insoluble and infusible. Because all the molecules are joined together, the molecular weight of a highly cross-linked polymer is a meaningless term since its value would depend upon the amount of material present.

The end groups of the polymer molecules are the functional groups that have not reacted at any stage. It is apparent that at exactly 100% conversion in a difunctional condensation, the reaction system would consist of only one molecule. The fact that the end groups of a condensation polymer can always undergo further reaction creates a difficulty in the high-temperature-melt spinning of polyamides and polyesters. To prevent subsequent changes in molecular weight, such as might occur in the melt during spinning at elevated temperatures, a monoacid or monoalcohol (molecular-weight modifier) is added to the original polymerization mixture. The excess of hydroxyl groups, for example, places a limit on the chain growth. That limit is reached when all the acid groups have reacted and all the end groups are hydroxyl. See CONDENSATION REACTION.

### ADDITION POLYMERIZATION

Unsaturated compounds such as olefins and dienes polymerize without the elimination of other products. The molecular weight and structure of the polymer are determined by the reaction conditions, that is, the nature of the catalyst or initiator, the temperature, and the concentration of reactants, monomer, initiator, and modifying agents. The unique feature of addition polymerization is the fact that the average chain length of the polymer formed initially is high and may increase fur-

ther through secondary branching reactions as the polymerization approaches completion.

The molecular-weight range for many useful addition polymers is relatively high, typically, from 20,000 to several million, as compared with the molecular weight range of 5,000–30,000 for typical condensation polymers.

The types of catalysis or initiation which are effective for addition polymerization may be identified in four groups: (1) free-radical catalysis by peroxides, persulfates, azo compounds, oxygen, and ultraviolet and other radiation; (2) acid catalysis by the Lewis acids, such as boron trifluoride, sulfuric acid, aluminum chloride, and other Friedel-Crafts agents; (3) basic catalysis, by alkali metals and metallic alkyls; and (4) heterogeneous catalysis, by chromic oxide on silica-alumina, nickel or cobalt on carbon black, molybdenum on alumina, and complexes of aluminum alkyls with titanium chloride.

The fourth group may be indeed a separate group; however, further information may show that it is simply a new example of one of the other groups or of some combination of them. It is convenient to discuss the mechanism and experimental methods of free-radical initiation as one subject, and to treat the remaining three types under the heading complex or ionic catalysts.

**Free-radical catalysis.** Among the several kinds of polymerization catalysis, free-radical initiation has been most thoroughly studied and is most widely employed. Atactic polymers are readily formed by free-radical polymerization of vinyl and diene monomers and some of their derivatives.

At an appropriate temperature, a peroxide decomposes to yield free radicals. In the presence of a monomer, the greater proportion of these radicals adds to the monomer and thereby initiates chain growth. The growing chains may terminate by coupling, by disproportionation, or by transfer with monomer, polymer, or added materials (transfer agents, retarders, and inhibitors). The series of equations on the facing page illustrates the reactions of initiation, propagation, and termination by coupling in vinyl acetate.

If transfer occurs with the unreacted monomer or polymer already formed, higher-molecular-weight branched structures will be produced, and if branching is excessive, insoluble products may be formed. If the radical produced in the transfer process is not sufficiently active to initiate a new chain, the transfer agent is called an inhibitor or a retarder. Mercaptans (RSH), carbon tetrachloride, and various organic solvents are examples of transfer agents, whereas amines and phenols are frequently used as inhibitors or retarders.

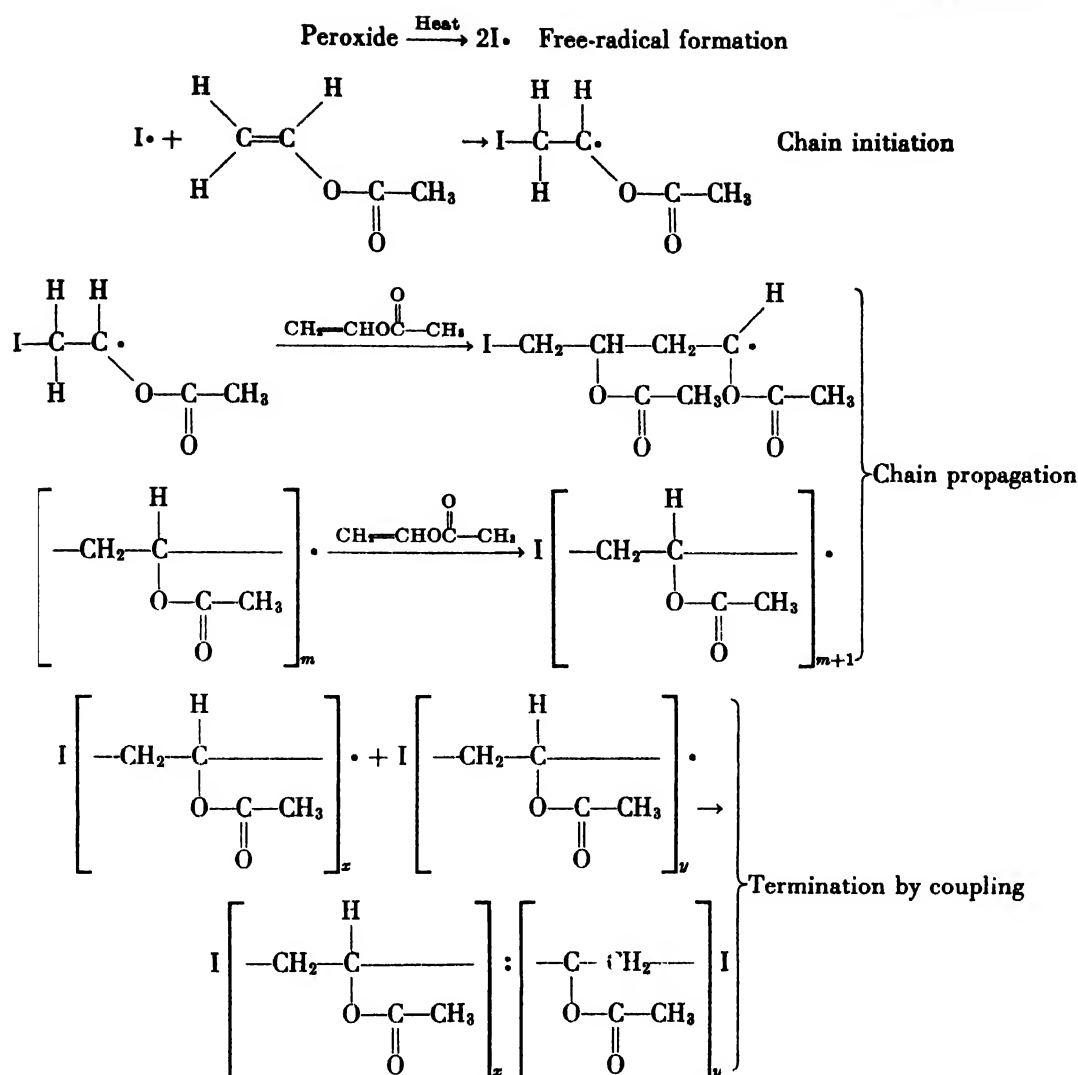
The rate of free-radical polymerization is increased by raising the temperature or increasing the concentration of monomer and initiator, whereas the molecular weight of the polymer is increased by increasing the monomer concentration, by lowering the temperature, and by lowering the concentration of initiator and transfer agents.

**Polymerization processes.** The bulk process consists of polymerization of the pure monomer in liquid form. On initiation by heat or light or very small amounts of azobisisobutyronitrile, a very pure polymer can be formed. The monomer and polymer are poor heat conductors; therefore the temperature of bulk polymerization is difficult to control. A further disadvantage is that small quantities of unreacted monomer are difficult to remove from the polymer. Polymerization in solution offers a means of carrying out the polymerization at lower monomer concentrations. Because solvents frequently act as transfer agents, polymerization in solution generally leads to the formation of lower-molecular-weight products.

Polymerization in aqueous emulsion has the advantages of giving a high rate of polymerization, a high molecular weight, and ease of temperature control. A liquid monomer is emulsified in water by use of a surface-active agent, such as soap. The soap micelles provide the polymerization centers. The free radicals (from a water-soluble initiator or growing chains of low molecular weight) diffuse into the soap micelle in which they react to form relatively linear polymer of high molecular weight. The polymer particles of small diameter, 500–1500 Å, are in stable suspension because the soap of the original micelle remains adsorbed in the outer layer of the polymer particle. The rate of emulsion polymerization and the molecular weight of the polymer increase with increasing numbers of micelle particles per unit volume. The product, a stable colloidal suspension of the polymer in water, usually is called a polymer latex or polymer emulsion. Polymer latexes are used directly for water-based paints, for adhesives, and for treating textiles. When polar solvents or electrolytes are added to the colloidal suspension, the polymer coagulates, and it can be separated and dried. In order to produce polymer emulsions which have the desired mechanical and thermal stability, it is frequently necessary to use moderately high concentrations of surface-active agents and protective colloids. Therefore, emulsion polymers are generally less pure than bulk polymers.

The redox initiation system was developed for polymerization in aqueous emulsions. In the presence of a water-soluble reducing agent such as sodium bisulfite or ferrous sulfate, the peroxide decomposes more rapidly at a given temperature, and consequently polymerization at useful rates can take place at lower temperatures. By the use of the redox system, the temperature for the commercial emulsion polymerization of styrene and butadiene is lowered from 50°C to 5°C to form "cold" rubber, which is higher in molecular weight than the "50°" or "hot" synthetic rubber.

The suspension polymerization system offers several advantages. By the use of a very small amount of surface-active materials and mechanical agitation, the monomer can be dispersed as droplets in the water. The monomer is not colloiddally dispersed but is temporarily broken up into droplets which

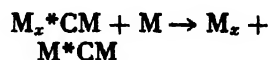
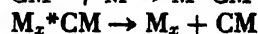
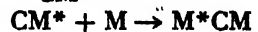
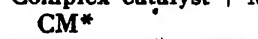


would coagulate if the stirring were discontinued. In the presence of a peroxide which is soluble in the monomer, bulk polymerization takes place in the droplets. When the polymerization reaches some 15–40%, the droplets containing the dissolved polymer become sticky and may coalesce. Various agents, such as talc and metallic oxides, have been recommended for use in very small amounts to prevent coagulation in the sticky stage. At higher degrees of conversion, the droplets will be transformed to hard balls of polymer containing dissolved monomer. At the completion of the polymerization, the balls or beads settle out. The beads may be dried easily and are ready for use. In the suspension system, the possibility of producing a pure polymer in bulk polymerization is combined with the ease of temperature control in the aqueous emulsion polymerization. There is the additional advantage that the product of suspension polymerization can be easily isolated for use.

**Complex or ionic catalysts.** Some polymerizations can be initiated by materials, often called ionic catalysts, that contain highly polar reactive sites or complexes. The term heterogeneous catalyst

is also applied to these materials because nearly all the catalyst systems are insoluble in monomers and other solvents. These polymerizations are usually carried out in solution from which the polymer can be obtained by evaporation of the solvent or by precipitation on the addition of a nonsolvent.

A general mechanism is shown in the following equations in which the growing chain is represented as an activated complex with the complex catalyst, without attempting to specify whether separate ions or free radicals are involved.



Initiative complex

Initiation

Termination by decomposition of complex

Termination by transfer to monomer

The distinguishing feature of complex catalysts is the ability of a few representatives of each type to initiate stereoregular polymerization or to cause the formation of polymers which can be crystallized. The polymerization process is visualized as

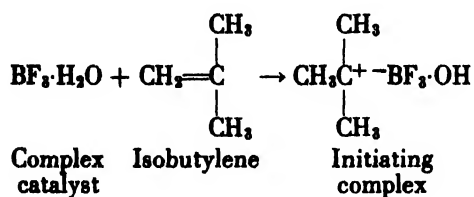


the formation of an activated complex of the monomer with the complex catalyst. For stereoregular growth to take place, the entering monomer must collide with the complex and be adsorbed in an oriented fashion. As reaction takes place, the new monomer assumes an activated condition within the complex catalyst and, at the same time, pushes the old monomer unit out. Chain growth is therefore similar to the growth of a hair from the skin.

The effect of conditions on rates of polymerization and on molecular size and structure is not yet fully understood. In general, the rate of polymerization is proportional to the concentrations of complex catalyst and monomer. The effect of temperature on the rate depends upon the stability and activity of the complex catalyst at the temperature under consideration. If the complex catalyst decomposes on increasing the temperature, then the rate of polymerization will be reduced. The effect of temperature upon molecular weight also depends upon the stability of the complex catalyst and upon the relative rates of propagation and termination. In some cases, at an optimum temperature of polymerization, the molecular weight depends upon the product of the ratio of the rate of propagation to termination and the monomer concentration, and in other cases, only upon that ratio of rates.

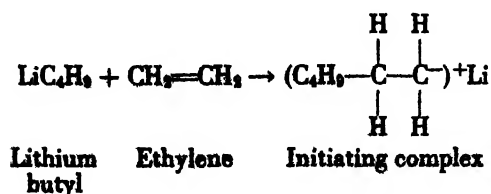
Examples of polymerization with the different types of complex catalysts are briefly described in the following paragraphs.

**Lewis acids.** Carbonium-ion catalysts such as  $\text{BF}_3$ ,  $\text{AlCl}_3$ , or  $\text{H}_2\text{SO}_4$  usually require the presence of a promoter such as  $\text{H}_2\text{O}$  or  $\text{HCl}$ .

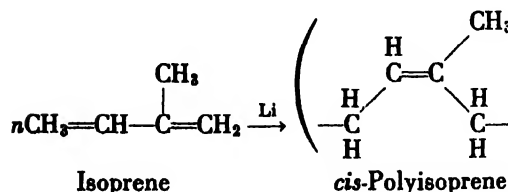
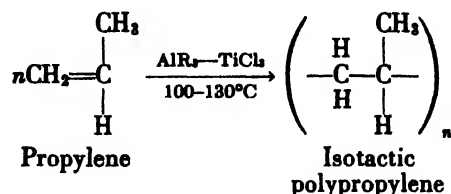


It has been suggested that the point of chain growth is a carbonium (positive) ion, although it is not contended that the ion necessarily has more than a transitory existence. Polymerization in the presence of Lewis acids takes place very rapidly at low temperatures,  $-100$  to  $-0^\circ\text{C}$ . The order of reactivity of some olefins in Lewis acid catalysis is vinyl ethers  $>$  isobutylene  $>$   $\alpha$ -methyl styrene  $>$  isoprene  $>$  styrene  $>$  butadiene.

**Alkali metals and metallic alkyls.** Carbanion catalysts such as sodium, lithium, and lithium butyl function at moderate temperatures,  $25$ – $150^\circ\text{C}$ . Inert hydrocarbon or ether solvents are generally used as reaction media.



It has been suggested that the point of chain growth is a negative or anionic center, although not necessarily a separate ion. The order of reactivity of some monomers in carbanion or anionic polymerization is acrylonitrile  $>$  methacrylonitrile  $>$  methyl methacrylate  $>$  styrene  $>$  butadiene. Heterogeneous catalysts (certain heavy metals or metal oxides on supports and complexes of aluminum alkyls with titanium chloride) function at moderate to high temperatures,  $50$ – $220^\circ\text{C}$ . Inert hydrocarbon or ether solvents are generally used as reaction media. The catalysts may be used in a fixed bed or as a slurry. Two examples are shown:



See ACID AND BASE; ADDITION REACTION; CATALYSIS; CHAIN REACTION, CHEMICAL; FIBER, MAN-MADE; FREE RADICAL; INHIBITOR (CHEMICAL); KINETICS (CHEMICAL); ORGANIC REACTION MECHANISM; OXIDATION-REDUCTION; PLASTICS FABRICATION; POLYMER; POLYMER, INORGANIC; POLYMER PROPERTIES. [J.A.M.; L.M.H.]

**Bibliography:** E. C. Bernhardt (ed.), *Processing of Thermoplastic Materials*, 1959; B. Golding, *Polymers and Resins*, 1959; H. Mark et al. (eds.), *High Polymers*, 12 vols., 1940–1958; *Modern Plastics Encyclopedia*, vol. 37, 1960; W. J. Roff, *Fibers, Plastics and Rubbers*, 1956; C. E. Schildknecht, *Vinyl and Related Polymers*, 1952; A. X. Schmidt and C. A. Marlies, *Principles of High-Polymer Theory and Practice*, 1948.

## Polymorphism (crystallography)

The property of crystallizing in two or more forms. The term is applied to crystals of the same substance having a different structure. Substances such as  $\text{CaCO}_3$  which exist in two crystal forms are said to be dimorphous, while substances such as  $\text{TiO}_2$  which appear in three forms are termed trimorphous. The polymorphic modifications of  $\text{CaCO}_3$  are aragonite, which is orthorhombic, and calcite, which is trigonal. The modifications of  $\text{TiO}_2$  are rutile (tetragonal), anatase (tetragonal), and brookite (orthorhombic). These modifications are stable at normal temperature and in a comparable temperature range. Other polymorphic forms, such as those of  $\text{SiO}_2$  (quartz, cristoballite, tridymite), have a specific nonoverlapping stability range.

p. W. Bridgman discovered many polymorphic modifications which are only stable under high pressure. The polymorphic forms of the elements are called allotropic modifications. See CRYSTAL STRUCTURE.

From a structural standpoint, molecular polarization and its change with temperature are important factors in bringing about changes from one structure to another. For substances having polymorphic forms stable at the same temperature, the atomic or ionic ratios are such that they are at the limit of the stability of a structure. Therefore the structure is sensitive to external secondary conditions such as temperature at which the crystals are formed, pressure, and impurities.

Polytypism is polymorphism in a narrow and specific sense. This term is applied to substances having structures like that of zinc blende. The zinc blende structure can be described as a close packing of sulfur layers, zinc layers in a similar arrangement being sandwiched in between. Calling *A, B, C* the three possible positions of sulfur layers and *A', B', C'* the zinc layers, the structure of zinc blende can be written symbolically as *AA' BB'CC'AA'BB'CC'*. A polytype is a structure in which longer sequences such as *ABC* are periodically repeated. Many polytypes of carborundum are known. The occurrence of such polytypes has been fully explained by the spiral growth of crystals. For a discussion of the chemical composition and crystal structure of polymorphous minerals, see MINERALOGY. [W.D.]

**Bibliography:** H. Baumhauer, *Z. Krist.*, 55:249, 1915; R. C. Evans, *Introduction to Crystal Chemistry*, 1939; F. C. Frank, Growth of carborundum: dislocation and polytypism, *Phil. Mag.*, 42(332): 1014-1021, 1951.

## Polymyxin

The basic polypeptide antibiotic produced as one of the fermentation products when certain strains of the bacterium, *Bacillus polymyxa*, are grown in suitable nutrients. Polymyxin's antibacterial action is uniquely directed only toward certain gram-negative bacteria. For examples of polypeptides active against gram-positive bacteria, see BACITRACIN; SUBTILIN.

Prior to public disclosure in 1947, developmental work on polymyxin had taken place independently in laboratories in the United States and Great Britain. The British investigators first published on polymyxin under the name "aerosporin," adopted from *Bacillus aerosporus*, a synonym for *B. polymyxa*. Detailed chemical studies on the purified antibiotic formed by each of a number of *B. polymyxa* strains revealed a family of closely related polymyxins. Furthermore, a single strain produced only one member of the family.

Only threonine and  $\alpha,\gamma$ -diamino butyric acid are common to all of the members of the family. The same  $C_{26}$  optically active fatty acid, present in all members, has been identified as *D*-6-methyloctan-1-*oic* acid.

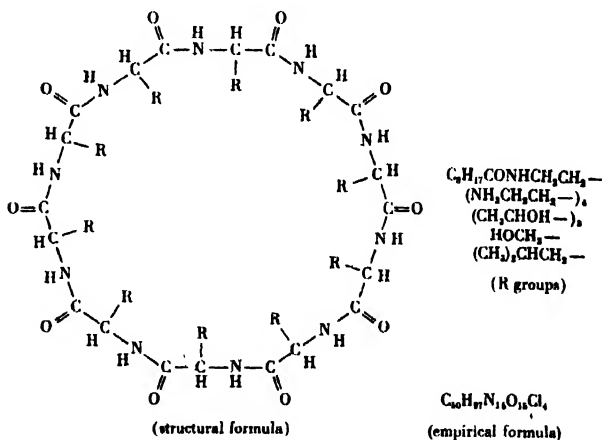
## Amino acid components of the polymyxins

Poly- myxin type	Leucine	Phenyl- alanine	Threo- nine	Serine	$\alpha,\gamma$ -diamino butyric acid
A	+	-	+	-	+
B	+	+	+	-	+
C	-	+	+	-	+
D	+	-	+	+	+
E†	+	-	+	-	+

\* Presence (+) or absence (-) of acid in molecule.

† Qualitative composition same as A but different ratio of constituent amino acids.

**Purification and chemical properties.** At the completion of 48-72 hours of fermentation time at 26-28°C, the antibiotic may be removed from the culture broth by adsorption on activated carbon, and then washed off the carbon by acid methanol. Precipitation in acetone is followed by extraction into butyl alcohol and, finally, conversion to polymyxin hydrochloride. The latter product is a nearly colorless, white powder, very soluble in water and methanol. It is insoluble in ethers, esters, ketones, hydrocarbons, and chlorinated solvents. The molecular weights range from 900-1150. The only polymyxin type for which a structural formula has been postulated is polymyxin D, in 1949. There is a hesitation to specify any arrangement of the groups on the ring structure because 5000 possibilities could exist.



Polymyxin D hydrochloride

**Toxicity.** The least toxic member of the family is polymyxin B. However, administration of type B to humans via the parenteral route is not advisable unless a serious infection by a polymyxin-susceptible bacterium has not responded to previous treatment by other chemotherapeutic drugs. The patient must be hospitalized so that possible toxic effects of polymyxin therapy, such as proteinuria or nitrogen retention, may be quickly ascertained. The drug has been formulated into ointments and troches, alone and with other antibiotics, and in this manner has been successfully used for topical treatment of various infections. [R.C.B.E.]

**Bibliography:** E. Jawetz, *Polymyxin, Neomycin, Bacitracin*, Antibiotics Monograph 5, 1956; P. H.



equations without system (4) having a solution. Nevertheless in many cases these equations make it possible to tell whether system (4) has solutions and to find them. The following example will illustrate the method.

$$\begin{aligned} f(x,y,z) &= x^2 - y + z = 0 \\ g(x,y,z) &= x + z^2 - y = 0 \\ h(x,y,z) &= x + z + 1 = 0 \end{aligned} \quad (5)$$

This gives  $R_x(f,g) = y^2 - 2z^2y - y + z^4 + z$ ,  $R_x(g,h) = y - z^2 + z + 1$ , and  $R_x(f,h) = -y + z^2 + 3z + 1$ . The necessary conditions are

$$\begin{aligned} R_y[R_x(f,g), R_x(g,h)] &= 2(2z + 1) = 0 \\ R_y[R_x(g,h), R_x(f,h)] &= 2(2z + 1) = 0 \end{aligned}$$

and

$$R_y[R_x(f,g), R_x(f,h)] = 4z(2z + 1) = 0$$

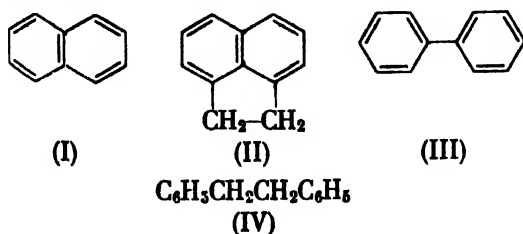
Now  $z = -\frac{1}{2}$  is the only value of  $z$  satisfying the necessary conditions. In system (5),  $z = -\frac{1}{2}$  gives  $x = -\frac{1}{2}$  and  $y = -\frac{1}{2}$ . Since these values satisfy all three equations, this is the unique solution of system (5).

Other methods of elimination are applicable when the equations of system (1) have a special form. If one of the equations is linear, as in system (5), such an equation can be solved for one of the variables and this variable can be eliminated. Systems which are linear (see LINEAR SYSTEMS OF EQUATIONS) in powers of the variables may be solved by the methods applicable to linear systems. See EQUATIONS, THEORY OF. [R.A.B.]

**Bibliography:** B. L. van der Waerden, *Modern Algebra*, vol. 1, rev. ed., 1953.

## Polynuclear hydrocarbon

One of a class of hydrocarbons possessing more than one ring. The aromatic polynuclear hydrocarbons may be divided into two groups. In the first, the rings are fused, which means that at least two carbon atoms are shared between adjacent rings. Examples are naphthalene (I), which has two 6-membered rings, and acenaphthene (II), which has two 6-membered rings and one 5-membered ring.



In the second group of polynuclear hydrocarbons, the aromatic rings are joined either directly, as in the case of biphenyl (III), or through a chain of one or more carbon atoms, as in 1,2-diphenylethane (IV).

Polynuclear hydrocarbons are found in the higher-boiling coal tar fractions, but some, especially those without condensed rings, for example, (III), and (IV), are most readily obtained from

benzene derivatives by synthesis. See ANTHRACENE; AROMATIC HYDROCARBON; BIPHENYL; DIPHENYLMETHANE; HEXAPHENYLETHANE; INDENE; NAPHTHALENE; PHENANTHRENE; STEROID; TRIPHENYLMETHANE. [C.K.B.]

## Polyolefin resins

Polymers derived from unsaturated hydrocarbons containing the ethylene or diene groups. Broadly, polyolefin resins may include virtually all addition polymers; however, the term polyolefin is specifically used for polymers of ethylene, the alkyl derivatives of ethylene (the  $\alpha$ -olefins), and for the dienes.

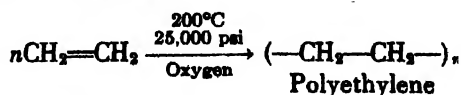
The polyvinyl resins, the fluorocarbon polymers, and other addition polymers are covered in other articles. This article includes discussions of the polymers of ethylene, propylene, and isobutylene, and brief mention is made of polymers of other  $\alpha$ -olefins and of butadiene, isoprene, and 2-chlorobutadiene. See POLYFLUOROOLEFIN RESIN; POLYVINYL RESINS.

**Polyethylene.** Polyethylene is a whitish, translucent polymer of moderate strength and high toughness. The available forms are partially crystalline. The physical properties vary markedly with the degree of crystallinity. The densities  $d$  of the products increase with increasing degrees of crystallinity, and it is common to classify the commercial grades as low density ( $d < 0.925$ ), medium density ( $d$  0.925–0.94), or high density ( $d > 0.94$ ). With increasing crystallinity or density, the products become stiffer and stronger, and have higher softening temperatures and higher resistance to penetration by liquids and gases; at the same time, they lose some of their resistance to tear, impact, and stress cracking, and higher temperatures and pressures are needed for molding.

Polyethylene is produced in very large volume. The major uses are as packaging films, containers, molded articles, electrical insulation, wire coating, and pipe.

Ethylene is produced on a large scale by the cracking of aliphatic hydrocarbons found in petroleum. The monomer can be conveniently produced in smaller volumes by the catalytic dehydration of ethanol.

The low- and medium-density polymers are formed by the polymerization of highly purified ethylene at about 150–250°C and 20,000–35,000 psi in the presence of a very small amount of oxygen or organic peroxide. At the higher temperatures, the low-density polymer is formed and at the lower reaction temperatures, the medium-density product is produced.



The high-density polymers are formed at relatively low temperatures and pressures (for example, 50–150°C and 100–2000 psi) in the presence of special catalysts, often referred to as stereospecific

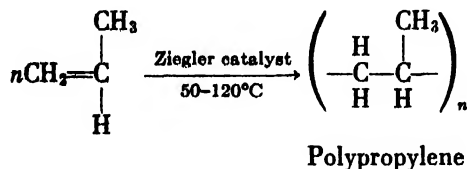
catalysts. These include oxidized forms of certain heavy metals such as chromium, reduced forms of heavy metals such as cobalt and nickel, and the Ziegler catalyst, a complex of an aluminum alkyl and a titanium chloride.

For the low-density material, the softening temperature and the maximum temperature for continuous use are about 105–115°C and 75°C, respectively; the corresponding temperatures for the high-density product are some 25–40°C higher.

Structural studies have shown that the higher-density polymers have highly linear structures and are approximately 85–95% crystalline. The lower-density materials are branched and are 50–85% crystalline. See ETHYLENE.

**Polypropylene.** High-molecular-weight, isotactic, highly crystalline polypropylene is generally similar in properties to high-density polyethylene. In comparison with the latter, isotactic polypropylene is harder and stronger, and softens at about 160°C.

Propylene is available in large quantities from the cracking of petroleum hydrocarbons, and the high-molecular-weight isotactic polymers are formed in the presence of the stereospecific catalyst used in the ethylene polymerization:

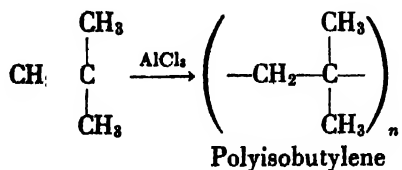


The crystalline product has been recently introduced, and should have many uses for molded objects, films, and fibers.

The low-molecular-weight polypropylene oils formed in the presence of acid catalysts, such as boron trifluoride or phosphoric acid, are useful in the manufacture of gasoline and synthetic detergents, but are not employed in plastics technology. See PROPYLENE.

**Polyisobutylene.** The polyisobutylene polymers vary in properties from low-molecular-weight oils to high-molecular-weight rubbery solids.

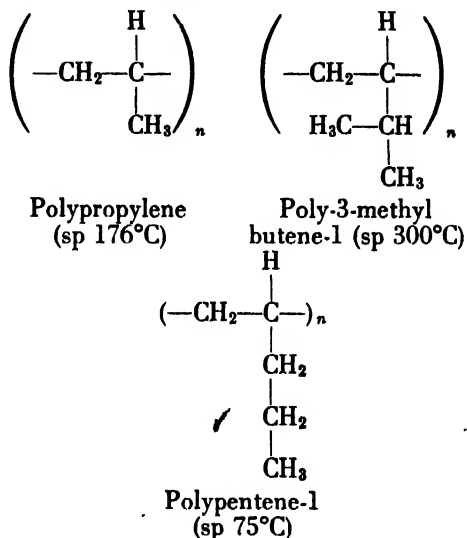
The monomer, obtained by cracking petroleum hydrocarbons, readily polymerizes in the presence of acid catalysts, such as boron trifluoride, aluminum chloride, or tin (IV) tetrachloride:



Polymerization conducted at 0–25°C yields oils which are useful in calking and sealing compositions. At low temperatures, such as –100 to –80°C, rubbery solids are formed. The solids are also useful in calking compositions and adhesive formulations; however, the main use of polyisobu-

tylene is in the form of the copolymer with 2–4% isoprene. The copolymer, known as butyl rubber, can be prepared at –90°C in the presence of methyl chloride as a diluent and aluminum chloride as the catalyst. The product, distinguished by its impermeability to gases and its resistance to aging, is used in automobile tires and tubes.

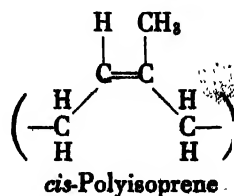
**Polymers of other 1-olefins.** The discovery of the stereospecific catalysts listed for ethylene polymerization has made possible the formation of high-molecular-weight, isotactic, crystalline polymers of other 1-olefins, such as 1-butene, 1-octene, and 1-dodecene. The softening temperatures (sp) of the crystalline, isotactic polymers of some of the 1-olefins are relatively high:



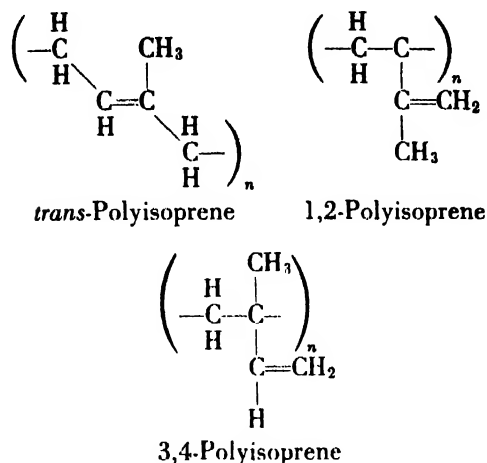
The high melting points of certain of the isotactic crystalline polymers may result from special spatial arrangements necessary to accommodate the presence of bulky branched groups near the polymeric chain. The high-melting polymers are of considerable interest because relatively few thermoplastic polymers are available which have high softening temperatures and at the same time can be easily fabricated.

**Polydienes.** Butadiene, isoprene, and 2-chlorobutadiene have gained the widest use of dienes employed in the polymer field. Butadiene and 2-chlorobutadiene have long been used in the production of synthetic rubbers by free-radical catalysis. The isoprene structure has long been recognized to correspond to the repeating unit in natural rubber. See RUBBER.

Special stereospecific catalysts are useful in the polymerization of butadiene and isoprene. Natural rubber consists largely of *cis*-polyisoprene,



and is characterized by high elasticity and low internal friction (on flexing). The polymers of butadiene and isoprene formed by free-radical catalysis contain mixtures of the *cis* and *trans* forms, together with some 1,2 or 3,4 structures, or a mixture of all four forms.



The presence of *trans* and 1,2 and 3,4 structures causes the rubber to have lower elasticity and higher internal friction on flexing.

The application of the stereospecific catalysts to the polymerization of isoprene and butadiene has led to the development of synthetic rubbers which contain high proportions of the *cis* structure and which are essentially equivalent in properties to natural rubber. See ADDITION REACTION; ALKENE; 1,3-BUTADIENE; DIENE; ISOPRENE; PLASTICS FABRICATION; POLYMERIZATION.

[J. A. MANSON; L. M. HOBBS]

## Polyoma virus

A virus which is capable of producing a number of tumors in newborn mice, rats, guinea pigs, hamsters, and rabbits. It was first described in 1955 by S. E. Stewart during studies on Gross's leukemia.

When inoculated into newborn mice, the following tumors, alone or in combination, may be produced: parotid tumors, mammary adenocarcinomas, thymomas, mesotheliomas, bone tumors, sweat gland carcinomas, and many other odd tumors. In newborn rats the same virus produces mostly sarcomas; in hamsters, angiomas and sarcomas; and in rabbits, a fibroma that regresses.

The virus grows readily in mouse embryo tissue culture and destroys the cells (cytopathogenic effect) with the production of a hemagglutinin. It is potentiated by growth in tissue culture, where it attains a high titer. It transforms hamster cells growing in tissue culture, which are then capable of producing sarcomas in hamsters. During the transformation process the infectious virus disappears, although an antigen to the virus can be demonstrated in the tumors.

The virus readily produces antibodies in injected animals, and they are often found in uninoculated



Electron micrographs of ultrathin sections of hamster kidney infected with the polyoma virus. (a) Two cells of a tumor in the hamster kidney infected with the polyoma virus. In the cytoplasm of both cells are inclusion bodies composed of polyoma virus particles. (b) Inclusion body shown on the left-hand side in (a), at higher magnification; characteristic polyoma virus particles may be seen. (c) Inclusion body, shown on the right-hand side in (a), at higher magnification; polyoma virus particles are present in an orderly array. (Leon Dmochewski, Clifford E. Gray, and Elizabeth Berezsky)



stock mice. It spreads in the laboratory, probably from the inoculated newborn, for virus is present in the urine. It is found in mice throughout the world but causes no disease, so its ability to produce tumors is strictly a laboratory phenomenon.

Antibodies can be measured by four types of *in vitro* tests: inhibition of cytopathogenicity, the mouse-antibody protection test, the complement-fixation test, and the hemagglutination-inhibition test. Of these, the last is most extensively used and has shown that infection occurs naturally and that the virus is carried in many transplantable mouse tumors, where it conveys immunity to the host. Humans have no natural antibodies to the virus.

[A. E. MOORE]

**Bibliography:** S. E. Stewart and B. E. Eddy, Tumor induction by SE polyoma virus and the inhibition of tumors by specific neutralizing antibodies, *Am. J. Public Health*, 49(11):1493-1496, 1959.

### Polyoxyethylation of alcohol

The process of effecting reaction of alcohols with ethylene oxide to produce polyethers. The polyethers so produced are characterized by a repeating chemical structure of oxyethyl groups. They have the formula,  $R-O-[CH_2-CH_2-O]_nH$ , in which  $n$  represents an integer from 1 to 20, and  $R$  represents the alkyl residue of the alcohol, ROH.

Although polyoxyethylation of alcohols is formally a polymerization reaction, the products are not of the extremely high molecular weight that characterizes the high polymers obtained, for example, by the polymerization of organic vinyl monomers. In fact, a special and useful class of compounds, in which  $n = 1$  in the above formulation, can be obtained by varying reaction conditions; this process is known as hydroxyethylation of alcohols. As a rule, however, the products are low-molecular-weight polyethers in the molecular weight range of 200-1000. See POLYMERIZATION.

Polyoxyethylation of alcohols is accomplished by heating the alcohol and ethylene oxide in the presence of a catalyst; certain compounds, both acidic and basic, are effective catalysts. The molecular weight and nature of the product are determined by the amount and kind of catalyst, time and temperature of reaction, and molar ratio of ethylene oxide to alcohol. The larger ratios of ethylene oxide to alcohol generally give the higher-molecular-weight polyethers. The reaction is limited to primary and secondary alcohols.

The polyether products are either viscous liquids or low-melting, waxy solids which are soluble in water and most polar organic solvents. They are useful as water-soluble lubricants, emulsifying agents, and formulators for cosmetics and ointments, adhesives, and paper coatings. Polyoxyethylation of methanol leads to commercial products marketed as Methoxy Polyethylene Glycols, and the commercial Polyethylene Glycols (Carbowax).

They are also similar in structure to poly(ethyl-

ene oxide), although the latter polymers are of much higher molecular weight and are prepared by a different process. See POLYETHYLENE GLYCOL.

[D. L. HEYWOOD]

### Polyplacophora

An order, also known as the Loricata, of the class Amphineura. These mollusks are commonly called chitons. The body is elliptical and the dorsal shell comprises eight calcareous plates which overlap posteriorly. A muscular girdle surrounds the plates. The plates are composed of two layers, the inner articulamentum and outer tegmentum. The latter contains ectodermal tissue, located in canals, which terminates in phototropic structures, the aesthetes. Gills, which resemble ctenidia, vary in number from 6 to 80 pairs.

Chitons are found in shallow coastal waters and range from the Ordovician to the Recent. Common examples are *Chiton*, *Tonicella*, *Chaetopleura*, *Craspedochilus*, and *Cryptochitin*. See AMPHINEURA.

[C. B. CURTIN]

### Polyploidy

The occurrence of related forms possessing chromosome numbers which are multiples of a basic number ( $n$ ), the haploid number. Forms having  $3n$  chromosomes are triploids;  $4n$ , tetraploids;  $5n$  pentaploids, and so on. Autopolyploids are forms derived by the multiplication of chromosomes from a single diploid organism. As a result the homologous chromosomes come from the same source. These are distinguished from allopolyploids which are forms derived from a hybrid between two diploid organisms. As a result, the homologous chromosomes come from different sources. About one-third of the species of vascular plants have originated at least partly by polyploidy, and as many more appear to have ancestries which involve ancient occurrences of polyploidy. The condition can be induced artificially with the drug colchicine and the production of polyploid individuals has become a valuable tool for plant breeding.

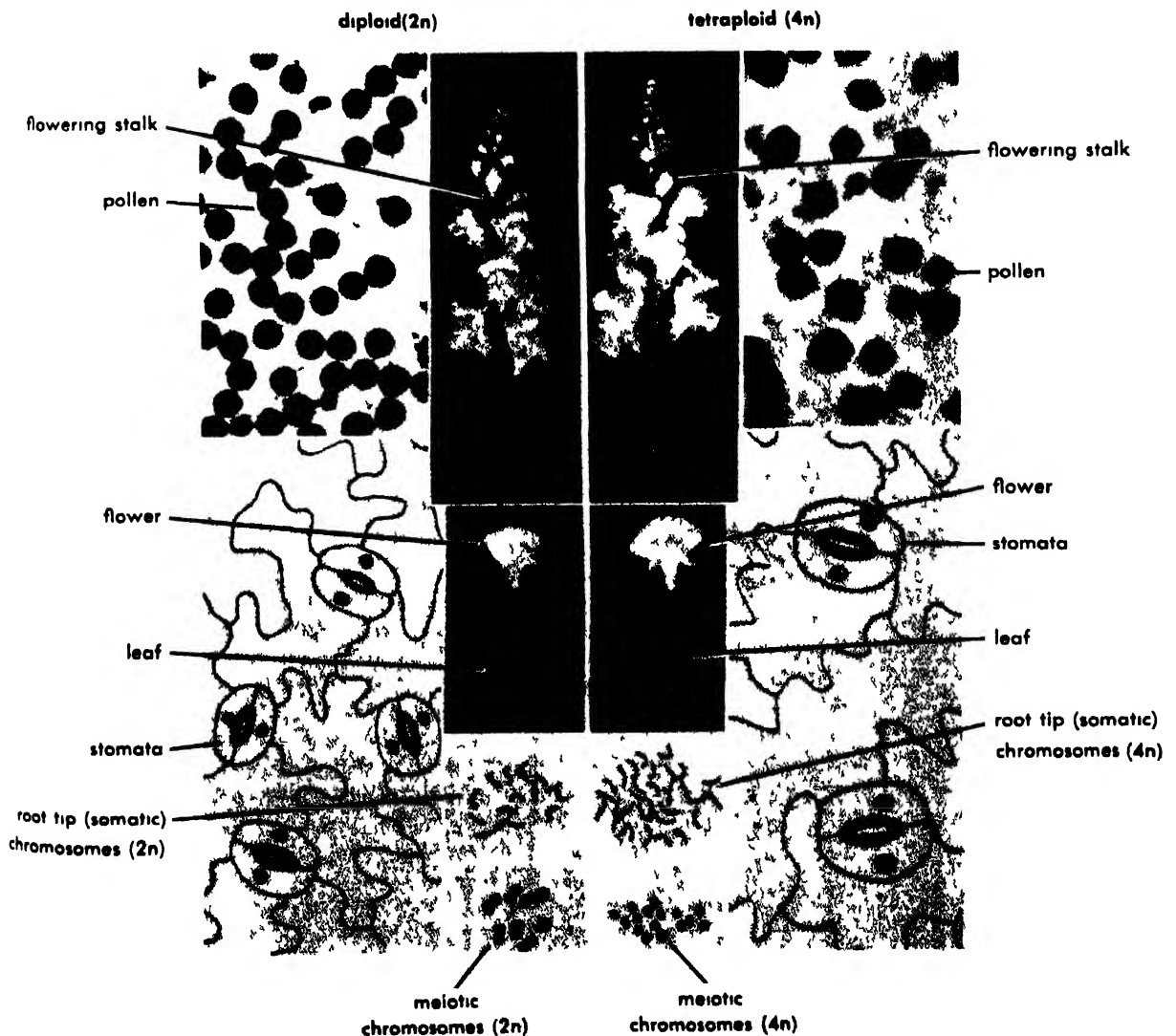
Polyploid series are irregularly distributed through the plant kingdom. They are particularly common in some families, notably the Gramineae and Rosaceae, and are rare in others, such as the Fagaceae and Umbelliferae. Not infrequently there exist related genera of the same family, one with and one without polyploidy, such as *Thalictrum* and *Aquilegia* of the family Ranunculaceae, or *Salix* and *Populus* of the family Salicaceae. Polyploidy has a significantly higher frequency in perennial herbs than in annual herbs and woody plants. A polyploid series often cited is that of the wheats in which the basic chromosome number is 7 and somatic chromosome numbers of 14, 28, and 42 occur.

In animals, undoubted examples of polyploidy are confined to groups which are parthenogenetic, such as crustaceans of the genus *Artemia*, certain

earthworms, weevils of the family Curculionidae, moths of the genus *Solenobia*, and sawflies of the genus *Diprion*; or which produce asexually by fission, as the flatworm *Dendrocoelum infernale*. A partial explanation of this situation is that in many animals the sex chromosome mechanism is so upset by polyploidy that sterile intersexes are produced. Because hybrid sterility in animals is usually genic rather than chromosomal in nature, and is not eliminated by chromosome doubling, allopolyploids can occur only rarely. Genic hybrid sterility, as in the mule, is the result of genes, contributed by the parents, interacting in the hybrid to disturb the course of meiosis and sex-cell formation. Chromosomal hybrid sterility is the result of the inability of homologous chromosomes to pair at meiosis due to rearrangement of the genes on the chromosome by a chromosomal aberration such as inversion, translocation, or deficiency.

**Hybridization.** Polyploidy and hybridization are usually associated with each other in evolution

When distantly related species are crossed, the sterile  $F_1$  hybrid often has little or no meiotic chromosome pairing. However in some cells the chromosome complement undergoes a doubling, leading to the formation of a few seeds from which second generation hybrids are formed. This fertile polyploid derivative possesses only bivalents, so that it breeds true for the intermediate condition. Such a polyploid is designated allopolyploid. Well-known examples are bread wheat (*Triticum aestivum*,  $2n = 42$ ), cultivated tobacco (*Nicotiana tabacum*,  $2n = 48$ ), and *Raphanobrassica* ( $2n = 36$ ) which is the hybrid between radish (*Raphanus sativus*,  $2n = 18$ ) and cabbage (*Brassica oleracea*,  $2n = 18$ ). Polyploids have also originated from hybrids between closely related species (*Primula kewensis*), or between subspecies of the same species (*Dactylis glomerata*). These are designated as segmental allopolyploids and autopolyploids of hybrid origin, respectively. Such polyploids may form varying numbers of trivalent and quadrivalent chromosome configurations and so may segregate



1. Diploid and tetraploid snapdragon.

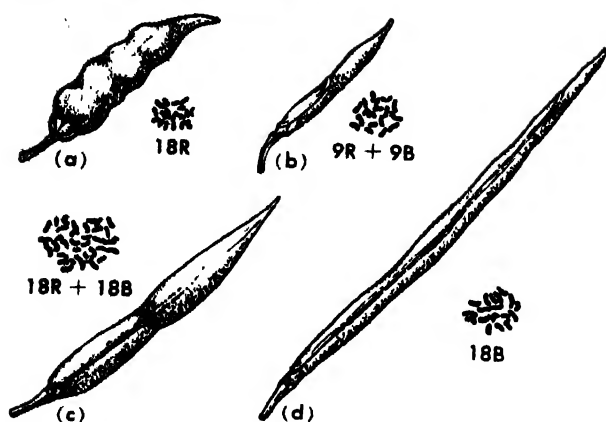


Fig. 2. Seed pods and somatic chromosome complements of radish, cabbage, and hybrids between them. (a) Radish (*Raphanus*). (b) The diploid hybrid of radish and cabbage. (c) Their allotetraploid hybrid *Raphanobrassica*. (d) Cabbage (*Brassica*). (After Karpechenko from E. W. Sinnott, L. C. Dunn, and T. Dobzhansky, *Principles of Genetics*, 5th ed., McGraw-Hill, 1958)

in the direction of their parental types. Autopolyploids derived from a single diploid race have been produced many times artificially but are rare in nature since they are usually weaker and less fertile than their diploid progenitor. In many plant genera, the simultaneous occurrence of several polyploids of these three types, plus others which combine characteristics of auto- and allopolyploidy, along with their diploid ancestors, has produced the polyploid complex. This consists of a series of usually distinct diploid species or subspecies, which represent extremes of morphological and ecological variation, plus a much larger number of polyploids, which form a network of connecting links between the diploids. Examples are *Bromus*, *Vaccinium*, *Galium*, and *Antennaria*.

**Distribution.** The geographic distribution of polyploids in relation to their diploid ancestors does not follow any consistent set of rules. Certain northern regions, such as Iceland and Spitsbergen, have floras with particularly high percentages of polyploidy, but so do some regions with subtropical or even tropical climates, such as New Zealand and Ceylon. Lowland and adjacent high alpine floras in western Europe do not differ significantly in percentage of polyploids. Diploid representatives of individual polyploid complexes tend to occupy geologically older habitats, and tetraploids are most prevalent in regions newly open to colonization. Consequently, the larger, older, and more stable land masses have relatively high percentages of diploids, while islands with floras derived through immigration, and areas disturbed by glaciation, volcanic activity, or other causes have high percentages of polyploidy.

**Plant breeding.** Artificial polyploids have been produced in most of the major species of crop plants by treating seeds or cuttings with colchicine. Generally, they have been less useful than their

diploid progenitors because of their slower growth and reduced fertility, but economically valuable autopolyploids have been produced in rye, sugar beets, rapeseed oil, red clover, snapdragons, marigolds, various orchids, and some other plants. Success has been achieved only when the doubling has been accompanied by intervarietal hybridization and selection. Although several artificial allopolyploids such as wheat-rye and wheat-*Agropyron* or quack grass have achieved partial success, none has yet been grown on a commercial scale. Artificial autopolyploids have also made possible the transfer of genes for disease resistance from wild species to cultivated species, even in instances in which the species are so distantly related that the  $F_1$  hybrids between their normal diploid forms are completely sterile. Examples are the transfer of rust resistance from goat grass (*Aegilops umbellulata*) to bread wheat and of resistance to both tobacco mosaic and black shank diseases from wild species of *Nicotiana* (*N. glutinosa*, *N. plumbaginifolia*) to cultivated tobacco. Polyploidy is, therefore, a useful tool in plant breeding when combined with hybridization and selection. See BREEDING (PLANT); CHROMOSOME ABERRATION; GENE; GENETICS; PLANT EVOLUTION; SPECIATION. [G.L.S.]

**Bibliography:** Brookhaven National Laboratory, *Genetics in Plant Breeding*, Brookhaven Symposia in Biol. 9, 1956; G. L. Stebbins, *Variation and Evolution in Plants*, 1950.

## Polypteriformes

A distinctive order of actinopterygian fishes, also called Cladistia, or the bichirs. Their characters include thick, rhombic, ganoid scales; a well-ossified internal skeleton; a symmetrical caudal fin, basi-



Bichir, *Polypterus endlicheri*; length to 3 ft. (After G. A. Boulenger, *Catalogue of the Fresh Water Fishes of Africa in the British Museum*, vol. 1, 1909)

cally heterocercal, with the upper part continuous with the dorsal fin; a dorsal series of free, spine-like finlets, each supported by a radial; a distinctive pectoral fin base with three enlarged radials; paired gular plates; paired ventral lungs; and a very large opisthotic.

This order consists of a single family, the Polypteridae, that is known from the Eocene (Cretaceous?). The two recent genera, *Polypterus* with about 10 species, and *Calamoichthys*, with 1 species, are confined to fresh waters of tropical Africa. Bichirs were long classified with Crossopterygians, but they are now believed to be modern descendants of early palaeonisciform actinopterygians. See ACTINOPTERYGII. [A.M.B.]

## Polysaccharide

A class of high-molecular-weight carbohydrates, colloidal complexes, which break down on hydrolysis to monosaccharides containing five or six carbon atoms. The polysaccharides are considered to be polymers in which monosaccharides have been glycosidically joined with the elimination of water. A polysaccharide consisting of hexose monosaccharide units may be represented by the following empirical equation:



The term polysaccharide is limited to those polymers which contain 10 or more monosaccharide residues. Polysaccharides such as starch, glycogen, and dextran consist of several thousand D-glucose units. Polymers of relatively low molecular weight, consisting of two to nine monosaccharide residues, are referred to as oligosaccharides. See DEXTRAN; GLUCOSE; GLYCOGEN; MONOSACCHARIDE; STARCH.

Polysaccharides are either insoluble in water or, when soluble, form colloidal solutions. They are mostly amorphous substances. However, x-ray analysis indicates that a few of them, such as cellulose and chitin, possess a definite crystalline structure. As a class, polysaccharides are nonfermentable and are nonreducing, except for a trace of reducing power due, presumably, to the free reducing group at the end of a chain. They are optically active, but do not exhibit mutarotation, and are relatively stable in alkali. See CELLULOSE; CHITIN; OPTICAL ACTIVITY.

The polysaccharides serve either as reserve nutrients (glycogen, inulin) or as skeletal materials (cellulose, chitin) from which relatively rigid mechanical structures are built. Some polysaccharides, such as certain galactans and mannans, however, serve both functions. Through the action of acids or certain enzymes, the polysaccharides may be degraded to their constituent monosaccharide units. Some polysaccharides yield only simple sugars on hydrolysis; others yield not only sugars but also various sugar derivatives, such as D-glucuronic acid or galacturonic acid (known generally as uronic acids), hexosamines, and even nonsugar compounds such as acetic acid and sulfuric acid.

The constituent units of the polysaccharide molecule are arranged in the form of a long chain, either unbranched as in cellulose and amylose, or branched as in amylopectin and glycogen. The linkage between the monosaccharide units is generally the 1,4- or 1,6-glycosidic bond with either the  $\alpha$  or  $\beta$  configuration, as the case may be. The branched glycogen and amylopectin contain both the 1,4 and 1,6 linkages. However, other types of linkage are known. In plant gum and mucilage polysaccharides, 1,2, 1,3, 1,5, and 1,6 linkages occur more commonly than the 1,4 type.

In an attempt to systematize the carbohydrate nomenclature, the generic name glycan was introduced as synonymous with the term polysaccha-

ride. This term is evolved from the generic word glucose, meaning a simple sugar, and the ending, "an," signifying a sugar polymer. Examples of established usage of the "an" ending are xylan for polymers of xylose, mannan for polymers of mannose, and galactomannan for galactose-mannose copolymers. Cellulose and starch are both glucans or glucoglycans, since they are composed of glucose units.

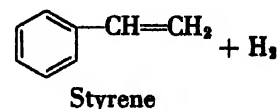
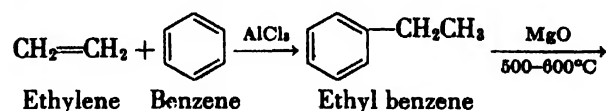
Polysaccharides are often classified on the basis of the number of monosaccharide types present in the molecule. Polysaccharides, such as cellulose or starch, that produce only one monosaccharide type (D-glucose) on complete hydrolysis are termed homopolysaccharides. On the other hand, polysaccharides, such as hyaluronic acid, which produce on hydrolysis more than one monosaccharide type (N-acetylglucosamine and D-glucuronic acid) are named heteropolysaccharides. See CARBOHYDRATE. [W.Z.H.]

*Bibliography:* R. L. Whistler and C. L. Smart, *Polysaccharide Chemistry*, 1953.

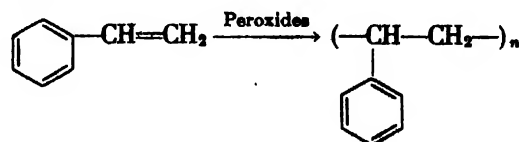
## Polystyrene resin

A hard, transparent, glasslike thermoplastic resin. Polystyrene is characterized by excellent electrical insulation properties, relatively high resistance to water, high refractive index, and low softening temperature.

Styrene is produced by the dehydrogenation of ethyl benzene, which, in turn, is obtained by the alkylation of benzene with ethylene.



Free-radical catalysts such as peroxides are often used for polymerization and copolymerization in bulk, solution, and in aqueous emulsion and suspension.



The high-molecular-weight homopolymers, copolymers, and polyblends are used as molding compounds for electronic mountings and insulation, toys, gift boxes, and panels. For a discussion of copolymers of styrene with unsaturated polyesters and with drying oils, see POLYESTER RESINS.

The copolymer of styrene and butadiene was the major synthetic rubber of World War II. During the 1940s, the redox system of polymerization was developed in which the presence of a reducing

agent caused the peroxide to yield free radicals more rapidly at lower temperatures. At the lower temperature of the redox polymerization, a more linear copolymer called cold rubber is obtained in high conversion with improved physical properties. Styrene-butadiene copolymers are still used in large volume for automobile tires and in various rubber articles. *See* POLYMERIZATION.

High-styrene-butadiene copolymers (containing more than 50% styrene) are resinous rather than rubbery. The latexes, as produced by emulsion polymerization, have achieved wide usage in water-based paints.

By sulfonation of the copolymer of styrene and divinyl benzene, an insoluble polyelectrolyte is produced. This product in the form of its sodium salt is employed as a cationic-exchange resin which is used for water softening.

The effects of blending small amounts of a rubbery polymer, such as butadiene-acrylonitrile rubber, with a hard, brittle polymer are most dramatic when the latter is polystyrene. The polyblend may have impact strength greater than ten times that of polystyrene.

The strength of amorphous, atactic polystyrene may be increased by cold drawing, even though crystallinity is not produced. Cold-drawn products in the form of filaments and sheets are available.

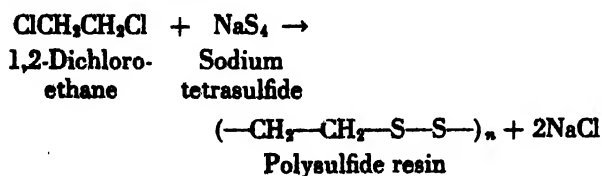
Crystallizable, isotactic polystyrene has been formed in the presence of (1) triphenylmethyl potassium in hexane solution, (2) the Alfin catalyst (sodium allyl-sodium isopropoxide) in hexane and benzene solutions, and (3) a Ziegler catalyst (titanium tetrachloride-aluminum triethyl) in petroleum ether solvent. The softening temperature of the crystalline polymer is substantially greater than that of the amorphous product.

Besides the many applications of styrene in combination with other materials as in rubber and paints, large quantities of the homopolymer and the polyblends are employed in the injection molding of toys, panels, novelty items, and also in the extrusion of sheets. The sheets are used for panels or they may be further shaped by vacuum forming for uses such as liners for refrigerator doors. *See* PLASTICS FABRICATION; RUBBER; STYRENE.

[J.A.M.; L.M.H.]

## Polysulfide resins

Resins that vary in properties from viscous liquids to rubberlike solids. Organic polysulfide resins are prepared by the condensation of organic dihalides with a polysulfide:



By the use of other dichlorides, such as bis(2-chloroethyl) ether,  $\text{ClCH}_2\text{CH}_2\text{OCH}_2\text{CH}_2\text{Cl}$ , the proper-

ties may be varied. The condensation is usually conducted in an aqueous medium from which the product may be separated and dried. Many of the polysulfide resins have an odor which is generally characteristic of monomeric sulfur compounds but is usually milder in nature.

The linear polymers can be cross-linked or cured by reaction with zinc oxide. Compounding and fabrication of the rubbery polymers can be handled on conventional rubber machinery. The polysulfide rubbers are distinguished by their resistance to solvents, such as gasoline, and to oxygen and ozone. The polymers are relatively impermeable to gases. The products are used to form chemically resistant coatings and special rubber articles, such as gasoline bags.

The polysulfide rubbers were among the very first commercial synthetic rubbers. Although the products are not as strong as other rubbers, their chemical resistance makes them useful in various applications.

The polysulfide rubbers were among the first polymers to be used in solid-fuel compositions for rockets. *See* ORGANOSULFUR COMPOUND; POLYMERIZATION; PROPELLANT; RUBBER.

[J.A.M.; L.M.H.]

*Bibliography:* H. Gilman (ed.), *Organic Chemistry*, vol. 1, 2d ed., 1943.

## Polytopes, regular

The  $n$ -dimensional analogs of the regular polygons ( $n = 2$ ) and platonic solids ( $n = 3$ ). They are conveniently denoted by their Schläfli symbols  $\{p, q, \dots\}$ ; for instance, the pentagon, hexagon, octagon, tetrahedron, octahedron are denoted by  $\{5\}$ ,  $\{6\}$ ,  $\{8\}$ ,  $\{3, 3\}$ ,  $\{3, 4\}$ . The cube is  $\{4, 3\}$  because its faces are squares  $\{4\}$  and there are 3 of them at each vertex. The five platonic solids  $\{p, q\}$ , which are the subject of Euclid's *Elements*, Book XIII, are determined by the inequality

$$(p - 2)(q - 2) < 4$$

The numbers of vertices, edges, faces ( $V, E, F$ ), as listed in the table, can be deduced from the obvious relations  $pF = 2E = qV$  with the help of Euler's formula  $V - E + F = 2$ .

The general polytope (sometimes loosely called a polyhedron regardless of the number of dimensions) is a finite region of  $n$ -dimensional space enclosed by a finite number of hyperplanes. When any redundant hyperplanes have been discarded, those that remain contain  $(n - 1)$ -dimensional polytopes called cells. For instance, the cells of a polygon are its sides, those of a polyhedron are its faces, and those of a 4-dimensional polytope are solids.

The platonic solid  $\{p, q\}$  is said to be regular because its faces are regular and its vertices are all surrounded alike. The 4-dimensional regular polytope  $\{p, q, r\}$  has 3-dimensional solid cells  $\{p, q\}$ ,  $r$  of which surround each edge; the 5-dimensional regular polytope  $\{p, q, r, s\}$  has cells  $\{p, q, r\}$ ,  $s$  of which surround each plane face; and so on.

Regular polytopes in  $n$  dimensions

Polytope	Schläfli symbol	Vertices	Edges	Faces	Solid cells	Hypersolid cells
$n = 2$						
$p$ -gon	$\{p\}$	$p$	$p$			
$n = 3$						
tetrahedron	$\{3,3\}$	4	6	4		
cube	$\{4,3\}$	8	12	6		
octahedron	$\{3,4\}$	6	12	8		
dodecahedron	$\{5,3\}$	20	30	12		
icosahedron	$\{3,5\}$	12	30	20		
$n = 4$						
5-cell	$\{3,3,3\}$	5	10	10	5	
8-cell	$\{4,3,3\}$	16	32	24	8	
16-cell	$\{3,3,4\}$	8	24	32	16	
24-cell	$\{3,4,3\}$	24	96	96	24	
120-cell	$\{5,3,3\}$	600	1200	720	120	
600-cell	$\{3,3,5\}$	120	720	1200	600	
$n > 4$						
simplex	$\{3,3,\dots,3\}$	$n + 1$	$\frac{1}{2}n(n + 1)$	....		$n + 1$
hypercube	$\{4,3,\dots,3\}$	$2^n$	$2^{n-1}n$	....		$2^n$
cross polytope	$\{3,\dots,3,4\}$	$2n$	$2n(n - 1)$	....		$2^n$

The six regular 4-dimensional polytopes  $\{p,q,r\}$  are determined by the inequality

$$p - \frac{4}{p} + 2q + r - \frac{4}{r} < 12$$

The 5-cell  $\{3,3,3\}$  may be drawn in perspective as a pentagon with its 5 diagonals, though in reality its 10 edges are all equal. The simplest drawing of

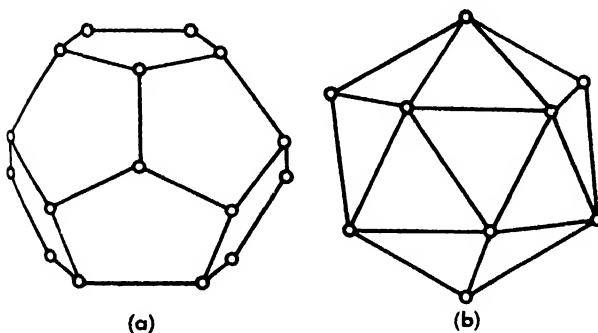


Fig. 1. Two regular polyhedrons. (a) Dodecahedron  $\{5,3\}$ . (b) Icosahedron  $\{3,5\}$ .

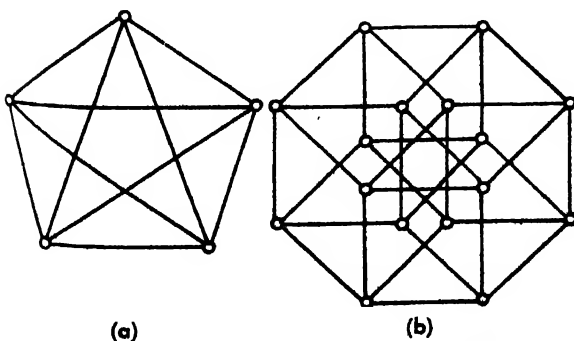


Fig. 2. Two regular 4-dimensional polytopes. (a) The 5-cell  $\{3,3,3\}$ . (b) The 8-cell  $\{4,3,3\}$ .

the 8-cell  $\{4,3,3\}$  consists of an octagon with a square placed inward on each side. The squares on two alternate sides are easily visualized as two opposite faces of a cube. The 8 such cubes are the cells of the 8-cell.

The regular tetrahedron can be inscribed in the cube, in the sense that the 4 vertices of the former occur among the 8 vertices of the latter. In the same sense, the cube can be inscribed in the dodecahedron, the 16-cell in the 8-cell, the 8-cell in the 24-cell, the 24-cell in the 600-cell, the 600-cell (and also the 5-cell) in the 120-cell.

The vertices of the  $n$ -dimensional simplex  $\{3,3,\dots,3\}$  consists of  $n + 1$  points, all equidistant from one another. Those of the cross polytope  $\{3,\dots,3,4\}$  are at equal distances from the origin in both directions along the  $n$  coordinate axes; thus their coordinates (for a cross polytope of edge  $\sqrt{2}$ ) are the permutations of  $(\pm 1, 0, \dots, 0)$ . Similarly, the  $2^n$  vertices of the hypercube  $\{4,3,\dots,3\}$ , of edge 2, are  $(\pm 1, \pm 1, \dots, \pm 1)$ . See ANALYTIC GEOMETRY; GEOMETRY, EUCLIDEAN. [H.S.M.C.]

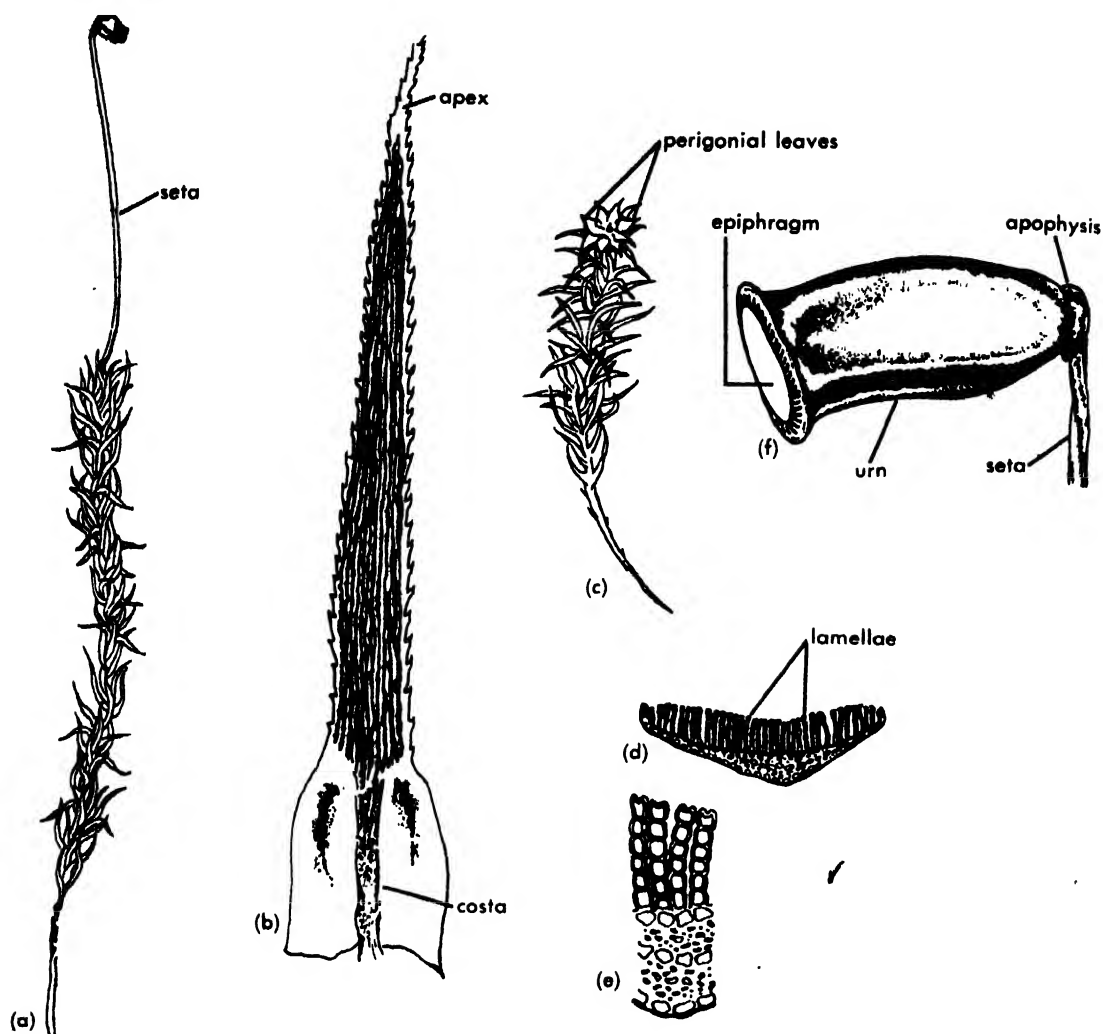
**Bibliography:** H. S. M. Coxeter, *Regular Polytopes*, 1948; D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, 1952; H. P. Manning, *Geometry of Four Dimensions*, reprint, 1955.

## Polytrichales

An order of acrocarpous, perennial mosses. They vary in size from small to large and short to tall. The stems are usually rigid, simple, or slightly branched. They arise from a prostrate subterranean rhizome. The stems are regarded as being highly developed, with a specialized central axis.

The leaves are mostly lanceolate from a sheathing base. The costa varies from narrow to broad and is often toothed on the lower surface. The upper leaf surface (both surfaces in *Oligotrichum*) has narrow, vertical, green lamellae attached by





*Polytrichum commune*. (a) Female plant. (b) Serrate apex of leaf. (c) Male plant. (d) Cross section of leaf showing lamellae. (e) Lamellae enlarged. (f) Urn of

*Polytrichum* sp. (From W. H. Welch, *Mosses of Indiana*, Ind. Dept. Conserv., 1957)

one edge to the costa and bistratose portion of the blade. The lamellae extend parallel to one another along the midrib, either few and distant or numerous and crowded. Each lamella is a few cells in height and one cell in width; the apical cell of the row often differs from the lower ones.

The inflorescence is usually dioecious. The male flower is terminal, large, and discoid. The female flower is terminal and budlike. The calyptra is cucullate, smooth, spinulose or spinulose-papillose, or with few to a felt of erect or deflexed hairs. The elongated seta bears a large capsule, which is oval, cylindric, or prismatic, with two to six angles. The hypophysis is fairly distinct in *Polytrichum*. The operculum is conic to convex and apiculate to rostrate. The peristome, regarded as a primitive type, is rarely lacking and is usually single. It consists of 16, 32, or 64 short, ligulate, unbarred teeth which are triangular in cross section. They arise from a basal membrane. The columella is

expanded at the apex into a shield-shaped membrane, the epiphragm, which covers the mouth of the urn and is united at its margin with the peristome teeth. See MUSCI; see also EUBRYA.

[W.H.W.]

### Polytropic process

An expansion (or contraction) of a gas during which some heat enters (or leaves) the system but not enough to maintain a constant temperature. During the polytropic expansion of a gas, external work is done both at the expense of some decrease in stored internal energy of the system and also at the expense of the heat transferred to the system from its surroundings. The polytropic path is compared to other thermodynamic processes on the accompanying graph. In comparison with the isentropic process, volume for volume, the polytropic path has a higher pressure and temperature, because the additional energy is provided by the

heat transfer process. See THERMODYNAMIC PROCESSES.

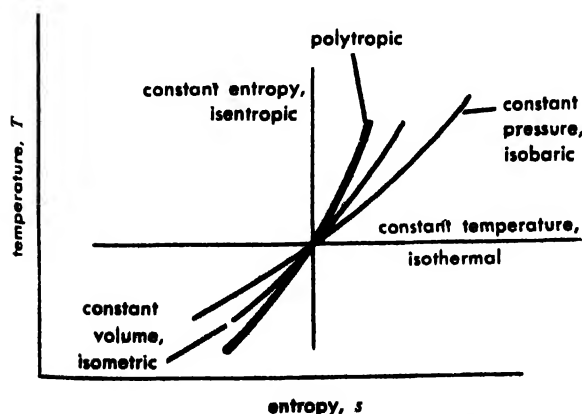
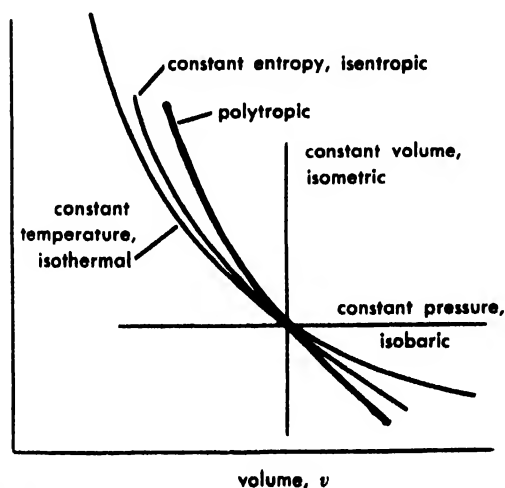
When a reversible polytropic expansion is plotted on  $P$ - $v$  coordinates, as illustrated, the result is a curve. However, if the expansion is plotted on log-log paper, the result is a straight line. The negative slope of this line equals the polytropic exponent,  $n$ , which characterizes the expansion. Thus

$$P_1 V_1^n = P_2 V_2^n = \text{constant}$$

Further, the work done during the expansion process is given by

$$W = \int_1^2 P dv = \text{constant} \int_1^2 V^{-n} dv = \frac{P_2 V_2 - P_1 V_1}{1 - n}$$

where  $W$  is work done in foot-lbs per pound of gas,  $P$  is absolute pressure in pounds per square foot,  $V$  is specific volume in cubic feet per pound, and  $n$



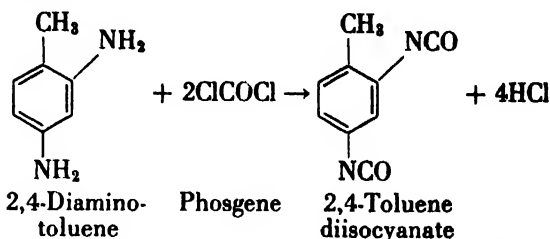
Polytropic process compared to other thermodynamic processes.

is constant for the process. For an expansion,  $n$  is less than the value of the isentropic  $k$  if heat is added to the gas while expanding, and it is greater than  $k$  if heat is transferred from the gas. See ISENTROPIC PROCESS. [J.B.]

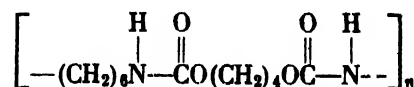
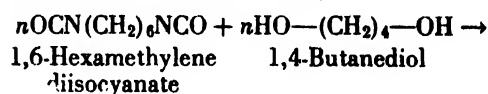
## Polyurethane resins

Resins that can be produced in forms varying from hard, glossy, solvent-resistant coatings, to abrasion- and solvent-resistant rubbers, and flexible or rigid foams. The foams have found the widest use in recent years. The flexible foams are employed as upholstery material for furniture, for rug backing, insulation, and crash pads. The rigid foams are employed as the core in structural laminates, such as in airplane wings.

Polyurethane (or polyisocyanate) resins are produced by the reaction of a diisocyanate with a compound containing at least two active hydrogen atoms, such as a diol, diamine, or dicarboxylic acid. Toluene diisocyanate and hexamethylene diisocyanate are frequently employed. They are prepared by the reaction of phosgene with the corresponding diamines.



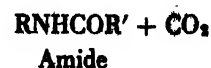
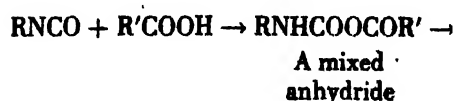
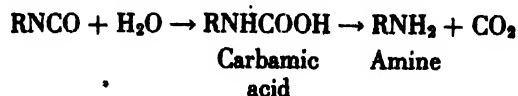
The condensation of a 1,6-hexamethylene diisocyanate with 1,4-butanediol,



A linear polyurethane

yields a linear polyurethane having excellent fiber-forming qualities. The fibers are generally similar to those of polyamides; however, they have lower softening points.

Polymers for coatings and foams are frequently prepared by the reaction of toluene diisocyanate with a polyester having unreacted  $\text{—OH}$  groups. For the production of foamed products, advantage is taken of the fact that the isocyanate group will react with water or carboxylic acids to yield carbon dioxide and either an amine or an amide:



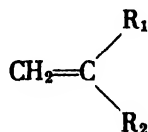
The amino groups can react with additional isocyanate to form cross-linkages. Thus by starting with a polyester having a known quantity of free carbox-

yllic groups as well as free —OH groups, or by adding a predetermined quantity of water to the polyester containing free —OH groups, insoluble cross-linked foams of desired density can be prepared. The foamed products can be cast in place around valves or between the walls of jacketed containers to provide mechanical or thermal insulation.

The flexible polyurethanes may be used for coating rubber articles to give them additional resistance to abrasion and solvents. Wire insulated with polyurethane resin can be soldered directly without previously removing the coating because the polymer decomposes at the soldering temperature to yield a clean wire surface. Among these various applications, the uses of the foamed products are developing most rapidly because of the ease of varying the density and flexibility, and the resistance to aging and solvents. See PLASTICS FABRICATION; POLYMERIZATION; URETHANE. [J.A.M.; L.M.H.]

## Polyvinyl resins

Polymeric materials generally considered to include polymers derived from monomers having the structure



in which  $\text{R}_1$  and  $\text{R}_2$  represent hydrogen, alkyl, halogen, or other groups. This article refers to polymers whose names include the term vinyl. Of these polymers, several have been used for a number of years, such as polyvinyl chloride, polyvinyl acetate, polyvinylidene chloride, polyvinyl alcohol, polyvinyl acetals, and polyvinyl ethers. Indeed, the terms vinyls and vinyl resins are frequently used to refer to the first three polymers of this group. Some polyvinyl resins of more recent origin are polyvinyl fluoride, polyvinylpyrrolidone, and polyvinylcarbazole. For discussions of other vinyl-type polymers, see POLYACRYLATE RESIN; POLYACRYLONITRILE RESIN; POLYFLUOROOLEFIN RESIN; POLYOLEFIN RESINS; POLYSTYRENE RESIN.

Many of the monomers can be prepared by addition of the appropriate compound to acetylene. For example, vinyl chloride, vinyl fluoride, vinyl acetate, and vinyl methyl ether may be formed by the reactions of acetylene with  $\text{HCl}$ ,  $\text{HF}$ ,  $\text{CH}_3\text{OOH}$ , and  $\text{CH}_3\text{OH}$ , respectively.

The polyvinyl resins may be characterized as a group of thermoplastics which, in many cases, are inexpensive and capable of being handled by solution, latex and injection molding, and extrusion techniques. The properties vary, depending upon chemical structure, crystallinity, and molecular weight.

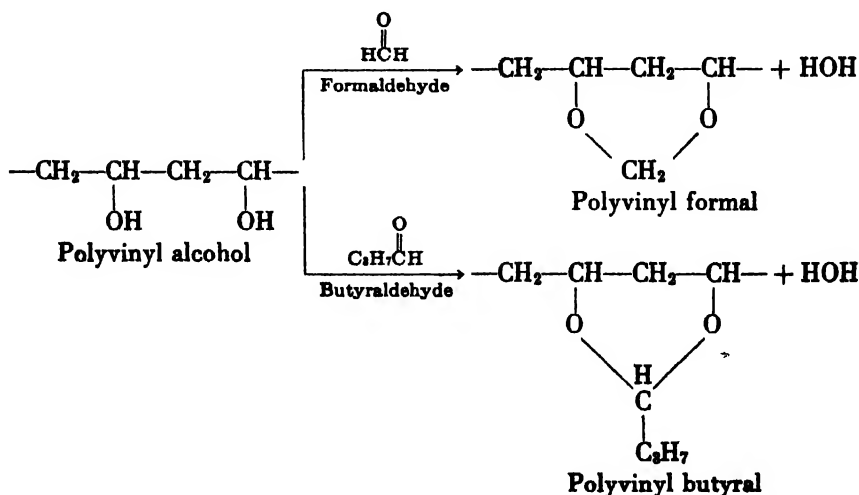
**Polyvinyl acetals.** These are relatively soft, water-insoluble thermoplastic products obtained by the reaction of polyvinyl alcohol with aldehydes. Polyvinyl butyral is soft and rubbery and is used primarily as the inner layer and binder for safety glass. Polyvinyl formal is the hardest of the group; it is used in making light-polarizing lenses and films, and to some extent in adhesive and wire-coating formulations.

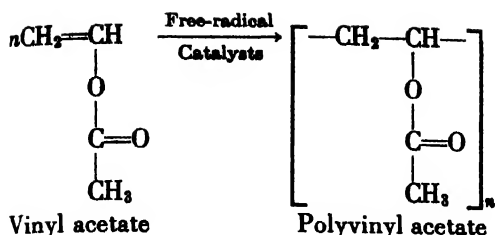
Polyvinyl butyral is usually obtained by the reaction of butyraldehyde with polyvinyl alcohol. The formal can be produced by the same process, but is more conveniently obtained by the reaction of formaldehyde with polyvinyl acetate in acetic acid solution.

**Polyvinyl acetate.** Polyvinyl acetate is a leathery, colorless thermoplastic material which softens at relatively low temperatures and which is relatively stable to light and oxygen. The polymers are clear and noncrystalline. The chief applications are as adhesives and binders for water-based or emulsion paints.

Polymerization and copolymerization may be conveniently effected by free-radical catalysis in aqueous emulsion and suspension systems. Vinyl acetate copolymerizes readily with various other vinyl monomers; however, it does not copolymerize with styrene by the free-radical process.

Anhydrous solid polymers and copolymers may be used directly in chewing gum and in adhesive formulations.





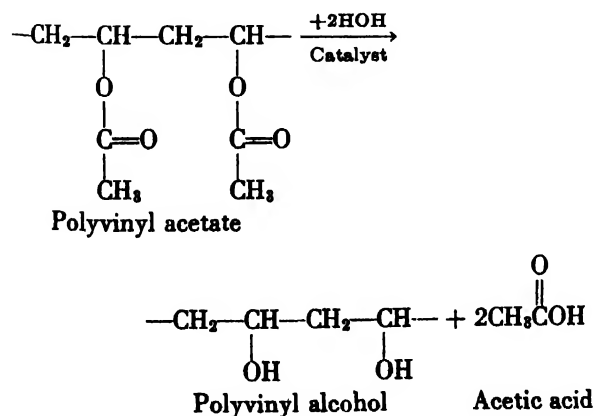
Aqueous dispersions, produced by the emulsion polymerization process and commonly called polymer emulsions or latices, are used for treating textiles and paper, as adhesives, and as water-based paints.

The water-based paints, prepared by pigmenting vinyl acetate polymer and copolymer emulsions, have achieved wide usage because of low cost of materials, ease of application, and resistance to weathering.

As water is removed from the latex by evaporation or absorption, the suspended polymer particles coalesce into a tough film. The character of the film may be modified by the use of comonomers in the original polymerization or by the addition of plasticizers to the final emulsions.

**Polyvinyl alcohol.** Polyvinyl alcohol is a tough, whitish polymer which can be formed into strong films, tubes, and fibers that are highly resistant to hydrocarbon solvents. Although polyvinyl alcohol is one of the few water-soluble polymers, it can be rendered insoluble in water by drawing or by the use of cross-linking agents.

So far, vinyl alcohol itself,  $\text{CH}_2=\text{CHOH}$ , has not been isolated; reactions designed to produce the monomer yield the tautomeric acetaldehyde instead. However, the polymer can be produced on a commercial scale by the hydrolysis of polyvinyl acetate.



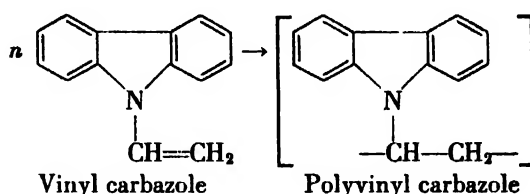
Two groups of products are available, those formed by the essentially complete hydrolysis (97% or greater) of polyvinyl acetate, and those formed by incomplete hydrolysis (50–90%).

The former, the "completely hydrolyzed" products, may be plasticized with water or glycols and molded or extruded into films and tubes and filaments which are resistant to hydrocarbons and can be rendered insoluble to water by cold-drawing or heat, or by the use of chemical cross-linking

agents. On cold-drawing, the degree of crystallinity is substantially increased. These products are used for liners in gasoline hoses, for grease-resistant coatings and paper adhesives, for treating paper and textiles, and as emulsifiers and thickeners. Insolubilized fibers have found large uses in Japan for clothing, industrial fabrics, and cordage.

The "partially hydrolyzed" products are generally more water-soluble and less subject to crystallization by drawing. These materials are used as emulsifying agents and thickeners, in steel-quenching solutions, in adhesive formulations, and in textile sizes.

**Polyvinyl carbazole.** Polyvinyl carbazole is a tough, glassy thermoplastic with excellent electrical properties and the relatively high softening temperature of 120–150°C. Polymerization can be carried out in bulk by free-radical catalysis. Uses



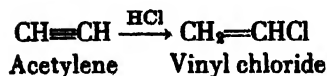
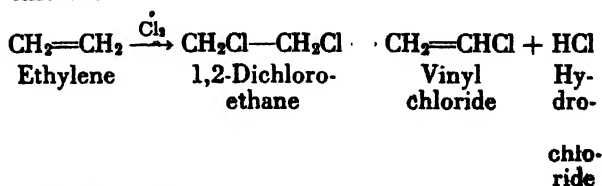
of the product are limited to special electrical applications requiring resistance to moderately high temperatures.

**Polyvinyl chloride.** Polyvinyl chloride is a tough, strong thermoplastic material which has an excellent combination of physical and electrical properties. The products are usually characterized as plasticized or rigid types.

The plasticized types, either soft copolymers or plasticized homopolymers, are elastic materials which are familiar in the form of shower curtains, floor coverings, raincoats, dishpans, dolls, bottle-top sealers, prosthetic forms, and packaging films, among others.

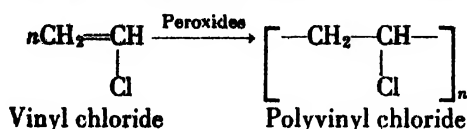
Rigid polyvinyl chloride products, which may consist of the homopolymer, copolymer, or polyblends, are commonly used in the manufacture of phonograph records, pipe, chemically resistant liners for chemical-reaction vessels, and for wire coating.

The monomer is frequently prepared from chlorine, acetylene, and ethylene by a combination of processes which affords complete utilization of the chlorine:



The polymerization of vinyl chloride and its copolymerization with other vinyl monomers may

be initiated by peroxides, and are conveniently



carried out in the presence of chain-transfer agents in aqueous emulsion or suspension systems.

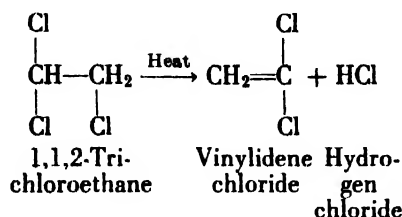
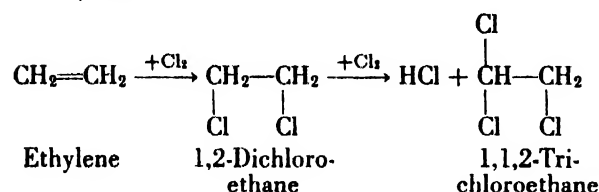
Because polyvinyl chloride products have a tendency to lose hydrogen chloride at high temperatures, a stabilizer such as calcium carbonate is usually included in the final composition.

Blends or alloys of polyvinyl chloride with small amounts of rubbery materials such as the copolymer of butadiene and acrylonitrile have been produced for applications such as panels and pipe in which impact resistance, as well as hardness and strength, is desired.

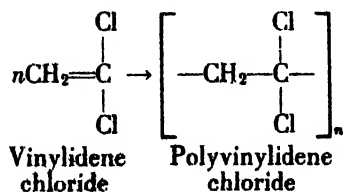
**Polyvinylidene chloride.** Polyvinylidene chloride is a tough, horny thermoplastic with properties generally similar to those of polyvinyl chloride. In comparison with the latter, polyvinylidene chloride is softer and less soluble; it softens and decomposes at lower temperatures, crystallizes more readily, and is more resistant to burning.

Because of its relatively low solubility and decomposition temperature, the material is most widely used in the form of copolymers with other vinyl monomers, such as vinyl chloride. The copolymers are employed as packaging film, rigid pipe, and as filaments for upholstery and window screens.

Vinylidene chloride is normally prepared by the pyrolysis of 1,1,2-trichloroethane. The latter is obtained by the chlorination of 1,2-dichloroethane which, in turn, is formed by the addition of chlorine to ethylene.



Polymerization as well as copolymerization may be initiated by peroxides and other free-radical



catalysts and is most satisfactorily effected by emulsion and suspension techniques.

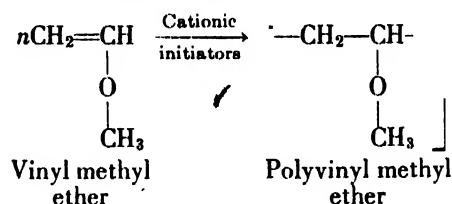
Because of the relatively low decomposition temperature of polyvinylidene chloride, a stabilizer such as an amine is normally included in the composition.

Films of polyvinylidene chloride, and especially the copolymer containing about 15% of vinyl chloride, are resistant to moisture and gases. Also, they can be heat sealed and have the property of shrinking on heating. By warming a food product wrapped loosely with a film of the polymer, a skin-tight, tough, resistant coating is produced.

By cold-drawing, the degree of crystallinity, strength, and chemical resistance of sheets, filaments, and even piping can be greatly increased.

**Polyvinyl ethers.** Polyvinyl ethers exist in several forms varying from soft, balsamlike semi-solids to tough, rubbery masses, all of which are readily soluble in organic solvents. Polymers of the alkyl vinyl ethers are used in adhesive formulations and as softening or flexibilizing agents for other polymers.

The monomers may be prepared by the reaction of alcohols with acetylene in the presence of alkali. Polymerization may be effected in bulk or solution at temperatures of  $-100$  to  $+25^\circ\text{C}$  by use of cationic initiators such as boron trifluoride. By

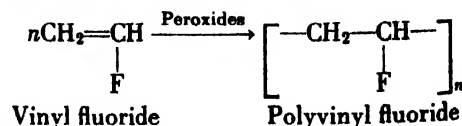


careful choice of conditions, it is possible to achieve stereoregular polymerizations which yield partially crystalline polymers that are harder and tougher than the amorphous products.

Polyvinyl methyl ether is soluble in cold water, but precipitates when the temperature is raised to about  $35^\circ\text{C}$ . The other alkyl vinyl ether polymers are insoluble in water.

**Polyvinyl fluoride.** Polyvinyl fluoride is a tough, partially crystalline thermoplastic material which has a higher softening temperature than polyvinyl chloride. Films and sheets are characterized by high resistance to impact and cracking caused by flexing and temperature, and to weathering.

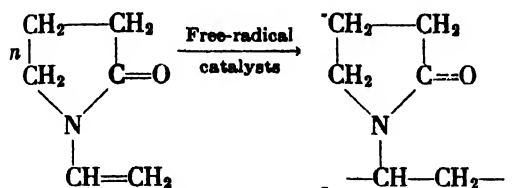
Polymerization can be effected in the presence of oxygen and peroxidic catalysts. Because of the low boiling point ( $-88^\circ\text{C}$ ) and high critical tempera-



ture of the monomer, polymerization is accomplished by use of pressure techniques similar to those employed in the high-pressure process for polymerizing ethylene. Like other polyvinyl halides, polyvinyl fluoride tends to lose the halogen acid at elevated temperatures.

Commercial production has just been initiated, and it seems likely that films and filaments of polyvinyl fluoride will find many applications.

**Polyvinyl pyrrolidone.** Polyvinyl pyrrolidone is a water-soluble polymer of basic nature which has film-forming properties, strong absorptive or complexing qualities for various reagents, and the ability to form water-soluble salts which are polyelectrolytes. The polymer can be prepared by free-



radical polymerization in bulk or aqueous solution. Isotonic solutions were used in Germany in World War II as an extender for blood plasma. The main current uses are as a water-solubilizing agent for medicinal agents such as iodine, and as a semi-permanent setting agent in hair sprays. Certain synthetic textile fibers containing small amounts of vinylpyrrolidone as a copolymer have improved affinity for dyes.

**New resins.** New polyvinyl resins are made available, often in the form of copolymers, when effective combinations of monomer preparation, polymerization methods, and property-use relationships are developed. For example, vinyl 2-ethylhexoate, vinyl oleate, and vinyl stearate are currently being used in copolymers; on the other hand, the homopolymer of vinyl stearate, a waxy substance, is employed in wax formulations and as a leather-treating agent. It is likely that new polyvinyl resins will be making their way into common usage over a period of many years. See PLASTICS FABRICATION; POLYMER PROPERTIES; POLYMERIZATION.

[J.A.M.; L.M.H.]

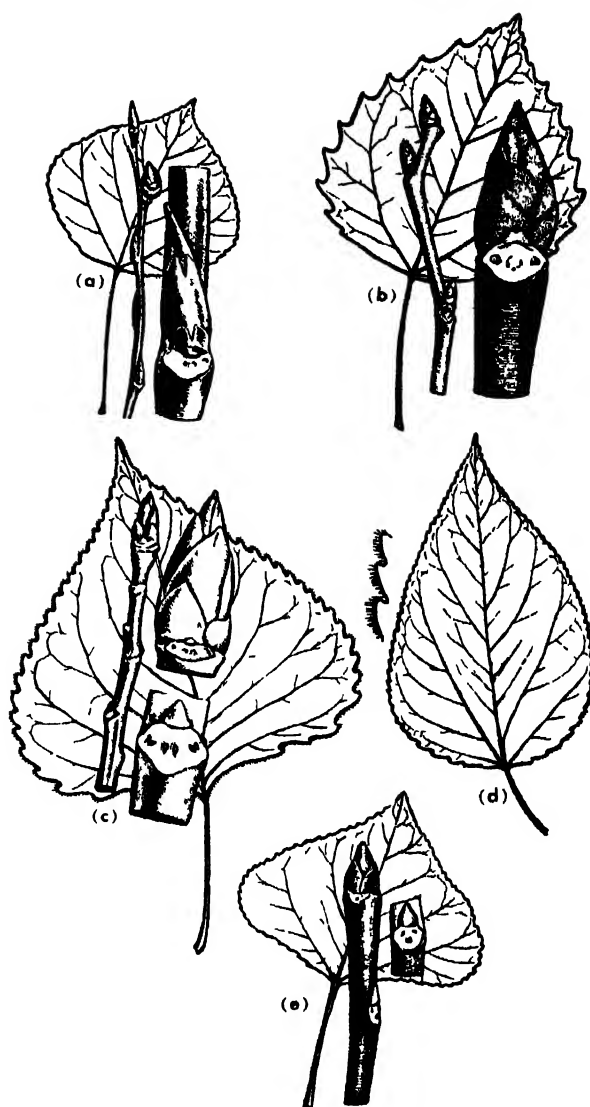
## Pomegranate

Small deciduous trees of the species *Punica granatum*, belonging to the plant order Myrtales. Pomegranate is grown as an ornamental as well as for its fruit. Propagation is by cuttings and occasionally by layering. The pomegranate is a native of Asia, having been described by writers as early as 300 B.C. It was originally known for its medicinal qualities, and cures for various ills were attributed to the fruit juice, the rind, and the bark of the roots. The fruit is a reddish, pomelike berry, containing numerous seeds imbedded in crimson pulp, from which an acid, reddish juice may be obtained. Limited quantities are grown in California and the Gulf States. See FRUIT (TREE); MYRTALES; STEM CUTTINGS.

[J.H.CE.]

## Poplar

Any tree of the genus *Populus*, family Salicaceae, marked by simple, alternate leaves which are usually broader than those of the willow, the other American representative of this family. Poplars have scaly buds, bitter bark, flowers and fruit in



(a) Quaking aspen or trembling aspen, *Populus tremuloides*. (b) Bigtooth aspen, *P. grandidentata*. (c) Cottonwood or necklace poplar, *P. deltoides*. (d) Balsam or tacamahac poplar, *P. balsamifera*. (e) Lombardy poplar, *P. nigra* var. *italica*.

catkins, and a 5-angled pith. See SALICALES; WILLOW.

Some species are commonly called cottonwood because of the cottony hairs attached to the seeds. Other species, called aspens, have weak, flattened leaf stalks which cause the leaves to flutter in the slightest breeze. One of the important species in the United States is the quaking aspen, *P. tremuloides*, which attains a height of 90 ft in the Rockies but is smaller in the East. This tree is widely distributed in North America from Labrador to Alaska. In the West it extends south through the Rockies and California to New Mexico and Lower California, and in the eastern United States it grows as far south as West Virginia. It is readily recognized by its comparatively small, finely toothed leaves, and by the shiny pointed winter buds. The soft wood of this species is used for pa-



per pulp; several million board feet are cut annually. *P. grandidentata*, the bigtooth aspen, attains a height of 60–70 ft, and occasionally, a diameter of 2 ft. The bigtooth aspen has a more restricted range in the northeastern quarter of the United States. The leaves are usually of larger size,  $2\frac{1}{2}$ –4 in. long and have larger teeth, hence the common name. The buds are plumper and somewhat downy.

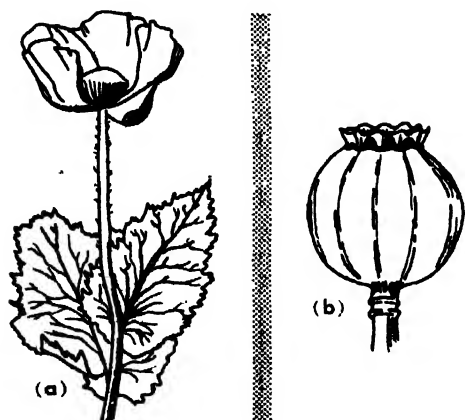
The European aspen, *P. nigra*, which is similar to the quaking aspen, is sometimes planted, and its variety, *italica*, the Lombardy poplar of erect columnar habit, is used in landscape planting.

The black cottonwood, *P. trichocarpa*, is the largest American poplar and is also the largest broad-leaved tree in the forests of the Pacific Northwest. It attains a height of 175–225 ft and a diameter of 7–8 ft. This tree ranges from southern Alaska to California and eastward through Washington and Oregon to Idaho, Montana, and Nebraska. The hairy fruit of the black cottonwood is a 3-valved capsule.

*Populus deltoides*, native in the eastern half of the United States, is a fast-growing tree which usually attains 80–100 ft in height and 3–4 ft in diameter, but under favorable conditions in the Mississippi Valley, it may attain a height of 150 ft and a diameter of 7–8 ft. The leaves are broadly triangular, hence the specific name, and the large terminal buds contain a pleasant-smelling balsamic resin. In *P. balsamifera*, the balsam or tacamahac poplar, the resin is used in medicine as an expectorant. The wood is used for veneer, boxes, crates, furniture, paper pulp, and excelsior. It is also planted as a shade tree and used in shelter belts. See FOREST AND FORESTRY; TREE. [A.H.G.]

## Poppy

A plant, *Papaver somniferum*, which is probably a native of Asia Minor. It is now cultivated extensively in China, India, and elsewhere. This plant is the source of opium obtained by cutting into the



(a) Opium poppy (*Papaver somniferum*). (From H. Kramer, *Applied and Economic Botany*, published by the author, 1914) (b) Capsule of poppy. (From W. E. Loomis and C. L. Wilson, *Botany*, rev. ed., Dryden, 1957)

fruits (capsules) soon after the petals have fallen. The white latex (juice) flows from the cuts and hardens when exposed to the air. This solidified latex is collected, shaped into balls or wafers, and often wrapped in the flower petals. This is the crude opium, which contains at least 20 alkaloids, including morphine and codeine. These drugs are used in medicine to allay pain, induce sleep, and relax spasms. Opium is one of the most useful drugs, but it is habit-forming and consequently should be used with the utmost caution. The opium habit is deleterious physically, mentally, and morally, and misuse of the drug is an extremely serious problem. See PAPAVERALES. [P.D.S.]

## Population dispersal

The process by which groups of living organisms expand the space or range within which they live. Because of their reproductive capacity, all populations have a natural tendency to expand. As increased area supports more individuals, dispersal and reproduction are intimately correlated.

Distinction should be made between dispersal and seasonal migration. Birds, butterflies, salmon, and others migrate regularly without necessarily expanding their geographic range, since they usually return to their original areas or die out.

**Dispersal phases.** Dispersal consists of several phases: (1) the production of units, that is, of individuals or parts of individuals (disseminules) fit or adapted for dispersal; (2) the transportation of individuals or disseminules to the new habitat; (3) ecesis, the process of becoming established through germination, rooting, physiological and psychological adjustment.

**Dispersal units.** These are disseminules (propagules, or diaspores) which may represent various stages of the life cycle of the individual. Many free-living animals do not produce special dispersal structures but rely upon the ability of the entire organism to move about (vagility). Organisms attached to a substratum, as most plants and certain animals, produce disseminules adapted to certain agents of dispersal. In order to be effective, a disseminule must have the ability to develop into one or more complete individuals. The structures listed in Table 1 are examples of disseminules. Sperm cells, unfertilized eggs and pollen grains, although capable of migration, are not true disseminules, because they cannot give rise to new individuals.

It is possible to analyze plant communities on the basis of morphological features of the disseminules. By assigning species to dispersal types one can construct dispersal spectra comparable to life form spectra in purpose and in usefulness.

**Transportation.** Individuals or disseminules are transported in five general ways: self-dispersal (autochory), water dispersal (hydrochory), wind dispersal (anemochory), animal dispersal (zoochory), and dispersal by man (anthropochory).

In active self-dispersal, autochory, the organism spreads in the course of its normal activities. The flight of starlings resulting in their gradual spread through the United States and the motility of the

Table 1. Examples of dispersal stages in life cycle of plants and animals

Environment	Organism	Disseminule	Dispersal by
Sea bottom	Kelp	Zoospore	Currents
	Coral	Planula	Currents
	Sea worm	Trochophore	Currents
	Clam	Trochophore	Currents
	Barnacle	Adult	Driftwood, ships
	Crab	Zoea	Currents
	Sea urchin	Pluteus	Currents
	Fish	Adult	Autochory
	Lamprey	Adult	Fish
	Mushroom	Spore	Wind
Terrestrial	Fern	Spore	Wind
	Pine	Seed	Wind
	Blueberry	Fruit	Birds
	Tumbleweed	Entire plant	Wind
	Insect	Adult	Autochory, wind
	Spider	Young animal	Wind
	Reptiles, birds, mammals	Adult	Autochory
Parasitic	Bacteria	Entire cell	Water, food, air
	Intestinal ameba	Cyst	Water, food, man
	Malaria parasite	Gamete, sporozoite	Mosquito
	Tapeworm	Egg	Pig
	Blood fluke	Egg, cercaria	Water, snail

Table 2. Plant dispersal types based upon morphological adaptations

Name	Definition	Example
Sarcchores	Disseminules fleshy	Cherry
Desmochores	Disseminules sticky or barbed	Cocklebur
Sporochores	Disseminules minute, light	Fern
Pogonochores	Disseminules plumed	Milkweed
Pterochores	Disseminules winged	Maple
Cyclochores	Spherical framework	Tumbleweed
Ballochores	Shot away by parent plant	Touch-me-not
Auxochores	Deposited by parent plant	Walking fern
Sclerochores	Disseminule without apparent adaptations	Violet
Barochores	Disseminules heavy	Oak

teria resulting in gradual spread through the nutrient media are examples. Certain plants possess mechanisms of self-dispersal as auxochores and ballochores listed in Table 2. In passive dispersal, one or more agents carry the dispersal unit to a new location. Such agents or vectors are water currents, wind, animals, any of man's vehicles such as trains, ships, and airplanes.

Water dispersal, or hydrochory, is prevalent in all marine and other aquatic populations. Plankton usually contains larval forms of bottom-dwellers (Fig. 1). Terrestrial forms associated with shore habitats are commonly dispersed by water. Buoyancy and resistance to salt water are a prerequisite for ocean dispersal. The first invaders of new islands such as Krakatau, are often of this type. Transoceanic similarities in floras and faunas have been explained partly by ocean currents.

Wind dispersal, anemochory, has various effects. It moves rolling disseminules in open deserts and grasslands (cyclochores, Fig. 2a); it deflects falling winged disseminules (pterochores, Fig. 2b, c, d), and it carries lightweight spores and dissemi-

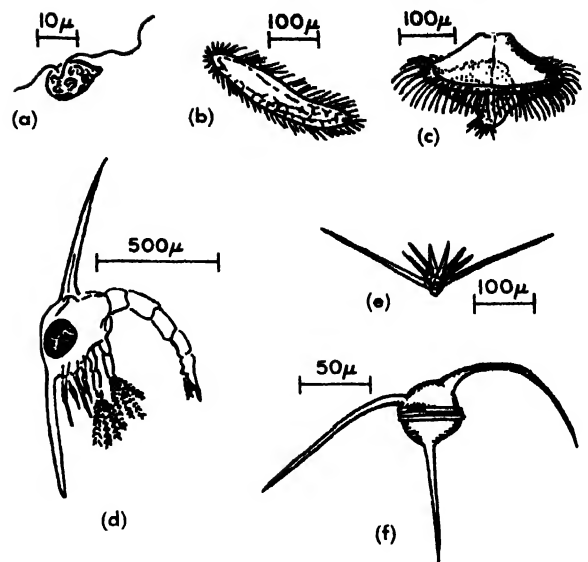


Fig. 1. Disseminules in plankton. (a) Kelp zoospore. (b) Coral planula. (c) Worm trochophore. (d) Crab zoea. (e) Brittle star pluteus. (f) Ceratium tripos.

nules with plumes for great distances (sporochores and pogonochores, Fig. 2e, h). Insects, spiders, and other light animals have been found many miles in the air, together with poplar seeds and other disseminules (Fig. 2f, g). Thus they may be carried hundreds of miles.

Animal dispersal, zoochory, is divided into epizoochory (barbed or sticky disseminules, desmochores, Fig. 3a, b, c) and endozoochory (disseminules eaten and egested by animals). Disseminules adapted to endozoochory are those like arillate seeds (Fig. 3d), common in the tropics and fruits with a fleshy mesocarp (Fig. 3f, g). Survival in the digestive tract of animals is a prerequisite. Bright fruit colors are frequent.

Dispersal by man, anthropochory, involves purposely dispersed organisms such as domesticated animals and plants and those accidentally transported such as weeds along railroads, beetles in grain shipments, birds, rats, barnacles, and starfish on and in ships.

*Ecesis.* Success in population dispersal depends upon three factors: fitness of the new habitat, fitness of the migrating individuals, and the chance of the juxtaposition of these two which, in the long run, depends on the number of individuals invading the new habitat. The probability for a new habitat to be favorable is greatest close to the parent population. Spores blown over great distances have less chance of landing in spots suited for germination than have seeds falling close to the parent plant. In wide-range dispersal larger numbers of disseminules are usually necessary than at close range, to insure ecesis.

The fitness of the individuals depends partly upon their genetic make-up. Offspring of organisms that reproduce without sexual union (apomictic) are likely to succeed only in identical habitats, as aphids, dandelions, and similar organisms. Off-

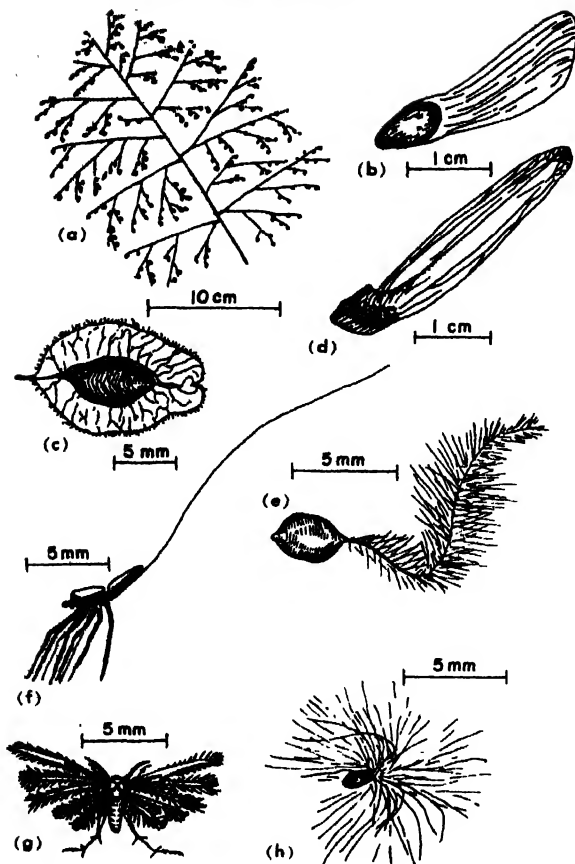


Fig. 2. Disseminules dispersed by wind. (a) Panic grass. (b) Pine seed. (c) Elm samara. (d) Tulip tree carpel. (e) Clematis carpel. (f) Spider. (g) Moth. (h) Cottonwood.

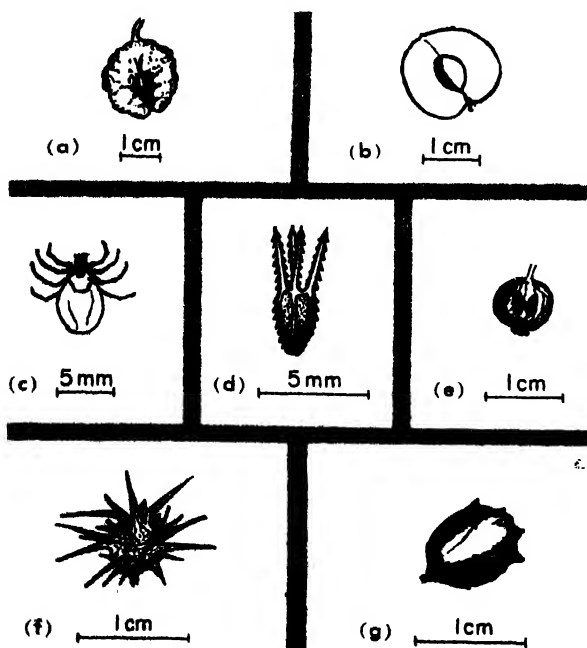


Fig. 3. Disseminules transported by animals. (a) Arillate legume seed. (b) Cherry (drupe). (c) Tick. (d) Beggar's tick fruit. (e) Currant berry. (f) Sandbur spikelet. (g) Juniper cone.

spring from self- or cross-fertilizing parents may succeed in a variety of situations. However, some hybrids which are sexually sterile are known to perpetuate themselves through apomixis. These are usually very successful locally.

**Barriers to dispersal.** A barrier is any discontinuity in the habitat greater than the maximum distance traveled by organisms in their normal dispersal. Oceans separating terrestrial habitats, continents separating marine habitats, mountain ranges intercepting wind dispersal, and deserts interrupting the continuity of forested land are all effective major barriers. Through the intervention of man these barriers are broken down in many cases. Since the development of frequent world travel thousands of species have become established on new continents as a result of anthropochory. See PARTHENOGENESIS; POPULATION DISPERSION; SPECIATION. [K.L.E.]

**Bibliography:** P. Dansereau and K. Lems. The grading of dispersal types in plant communities and their ecological significance, *Contrib. inst. botan. univ. Montréal*, 71, 1957; P. A. Fryxell, Mode of reproduction of higher plants, *Botan. Rev.* 23:135-233, 1957; P. A. Glick. *The Distribution of Insects, Spiders and Mites in the Air*, USDA Tech. Bull. 673, 1939; R. Hesse, W. C. Allee, and K. P. Schmidt, *Ecological Animal Geography*, 1937; E. J. Salisbury, *The Reproductive Capacity of Plants*, 1942.

## Population dispersal

The spatial distribution at any particular moment of the individuals of a species of plant or animal. Under natural conditions organisms are distributed either by active movements, or migrations, or by passive transport by wind, water, or other organisms. The act or process of dissemination is usually termed dispersal (see POPULATION DISPERSAL), while the resulting pattern of distribution is best referred to as dispersion. Dispersion is a basic characteristic of populations, controlling various features of their structure and organization. It determines population density, that is, the number of individuals per unit of area, or volume, and its reciprocal relationship, mean area, or the average area per individual. It also determines the frequency, or chance of encountering one or more individuals of the population in a particular sample unit of area, or volume. The ecologist therefore studies not only the fluctuations in numbers of individuals in a population but also the changes in their distribution in space.

**Principal types of dispersal.** The dispersion pattern of individuals in a population may conform to any one of several broad types, such as random, uniform, or contagious (clumped). Any pattern is relative to the space being examined; a population may appear clumped, when a large area is considered, but may prove to be distributed at random with respect to a much smaller area.

**Random or haphazard dispersal.** This implies that the individuals have been distributed by chance. In such a distribution, the probability of finding an individual at any point in the area

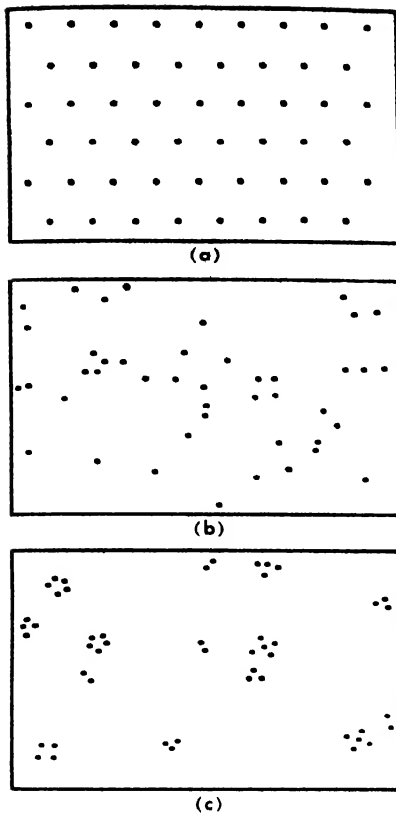


Fig 1 Three basic patterns of the dispersion of individuals in a population. (a) Uniform. (b) Random. (c) Clumped, but groups random. (E. P. Odum, *Fundamentals of Ecology*, Saunders, 1953)

same for all points (Fig. 1b). Hence a truly random pattern will develop only if each individual has had an equal and independent opportunity to establish itself at any given point. In a randomly dispersed population, the relationship between frequency and density can be expressed by the formula

$$F = 100(1 - e^{-D})$$

where  $F$  is percentage frequency,  $D$  is density, and  $e$  is the base of natural or Napierian logarithms. Thus when a series of randomly selected samples is taken from a population whose individuals are dispersed at random, the numbers of samples containing 0, 1, 2, 3, . . . ,  $n$  individuals conform to the well-known Poisson distribution

$$e^{-D}, De^{-D}, \frac{D^2}{2!} e^{-D}, \frac{D^3}{3!} e^{-D}, \dots, \frac{D^n}{n!} e^{-D}$$

Randomly dispersed populations have the further characteristic that their density, on a plane surface, is related to the distance between individuals within the population, in the following way

$$D = \frac{1}{4\bar{r}^2}$$

where  $\bar{r}$  is the mean distance between an individual and its nearest neighbor. These mathematical properties of random distributions provide the principal basis for a quantitative study of population dispersion (see *BIOMETRY*). Examples of approxi-

mately random dispersions can be found in the patterns of settlement by free-floating marine larvae and of colonization of bare ground by airborne disseminules of plants. Nevertheless, true randomness appears to be relatively rare in nature, and the majority of populations depart from it either in the direction of uniform spacing of individuals or more often in the direction of aggregation.

**Uniform dispersion.** This type of distribution implies a regularity of distance between and among the individuals of a population (Fig. 1a). Perfect uniformity exists when the distance from one individual to its nearest neighbor is the same for all individuals. This is achieved, on a plane surface, only when the individuals are arranged in a hexagonal pattern. Patterns approaching uniformity are most obvious in the dispersion of orchard trees and in other artificial plantings, but the tendency to a regular distribution is also found in nature, as for example in the relatively even spacing of trees in forest canopies, the arrangement of shrubs in deserts, and the distribution of territorial animals.

**Contagious or clumped dispersion.** The most frequent type of distribution encountered is contagious or clumped (Fig. 1c), indicating the existence of aggregations or groups in the population. Clusters and clones of plants, and families, flocks, and herds of animals are common phenomena. The degree of aggregation may range from loosely connected groups of two or three individuals to a large compact swarm composed of all the members of the local population. Furthermore, the formation of groups introduces a higher order of complexity in the dispersion pattern, since the several aggregations may themselves be distributed at random evenly, or in clumps. An adequate description of dispersion, therefore, must include not only the determination of the type of distribution, but also an assessment of the extent of aggregation if the latter is present.

**Analysis of dispersion.** If the type or degree of dispersion is not sufficiently evident upon inspection, it can frequently be ascertained by use of sampling techniques. These are often based on counts of individuals in sample plots or quadrats. Departure from randomness can usually be demonstrated by taking a series of quadrats and testing the numbers of individuals found therein for their conformity to the calculated Poisson distribution which has been described above. The observed values can be compared with the calculated ones by a chi-square test for goodness of fit, and lack of agreement is an indication of nonrandom distribution. If the numbers of quadrats containing zero or few individuals, and of those with many individuals are greater than expected, the population is clumped; if these values are less than expected, a tendency towards uniformity is indicated. Another measure of departure from randomness is provided by the variance-mean ratio, which is 1.00 in the case of the Poisson (random) distribution. If the ratio of variance to mean is less than 1.00, a regular dispersion is indicated; if the ratio is greater than 1.00, the dispersion is clumped.

In the case of obviously aggregated populations, quadrat data have been tested for their conformity to a number of other dispersion models, such as Neyman's contagious, Thomas' double Poisson, and the negative binomial distributions. However, the results of all procedures based on counts of individuals in quadrats depend upon the size of the quadrat employed. Many nonrandom distributions will seem to be random if sampled with very small or very large quadrats, but will appear clumped if quadrats of medium size are used. Therefore the employment of more than one size of quadrat is recommended.

The fact that plot size may influence the results of quadrat analysis has led to the development of a number of techniques based on plotless sampling. These commonly involve measurement of the distance between a randomly selected individual and its nearest neighbor, or between a randomly selected point and the closest individual. At least four different procedures have been used (Fig. 2). The closest-individual method (Fig. 2a) measures the distance from each sampling point to the nearest individual. The nearest-neighbor method (Fig. 2b) measures the distance from each individual to its nearest neighbor. The random-pairs method (Fig. 2c) establishes a base line from each sampling point to the nearest individual, and erects a 90° exclusion angle to either side of this line. The distance from the nearest individual lying outside the exclusion angle to the individual used in the base line is then measured. The point-centered quarter method (Fig. 2d) measures the distance from each sampling point to the nearest individual in each quadrant.

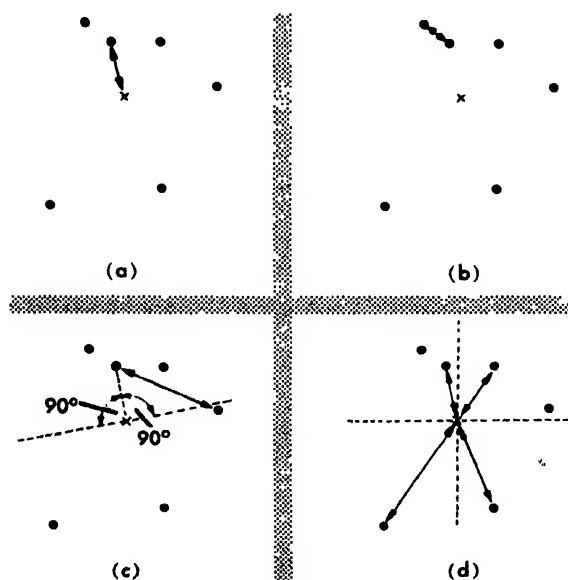


Fig. 2. Distances measured in four methods of plotless sampling. (a) Closest individual. (b) Nearest neighbor. (c) Random pairs, with 180° exclusion angle. (d) Point-centered quarter. x is the sampling point in each case. (P. Greig-Smith, *Quantitative Plant Ecology*, Butterworths, 1957)

In each of these four methods of plotless sampling, a series of measurements is taken which can be used as a basis for evaluating the pattern of dispersion. In the case of the closest-individual and the nearest-neighbor methods, a population whose members are distributed at random will yield a mean distance value that can be calculated by use of the density-distance equation which has been given above. In an aggregated distribution, the mean observed distance will be less than the one calculated on the assumption of randomness; in a uniform distribution it will be greater. Thus the ratio  $\bar{r}_A / \bar{r}_R$ , where  $\bar{r}_A$  is the actual mean distance obtained from the measured population and  $\bar{r}_R$  is the mean distance expected under random conditions, affords a measure of the degree of deviation from randomness.

Additional information about the spatial relations in a population can be secured by extending these procedures to measurement of the distance to the second and successive nearest neighbors, or by increasing the number of sectors about any chosen sampling point. However, since all of these methods assume that the individuals are small enough to be treated mathematically as points, they become less accurate when the individuals cover considerable space.

**Factors affecting dispersion.** The principal factors that determine patterns of population dispersion include (1) the action of environmental agencies of transport, (2) the distribution of soil types and other physical features of the habitat, (3) the influence of temporal changes in weather and climate, (4) the behavior pattern of the population in regard to reproductive processes and dispersal of the young, (5) the intensity of intra- and interspecific competition, and (6) the various social and antisocial forces that may develop among the members of the population. Although in certain cases the dispersion pattern may be due to the overriding effects of one factor, in general populations are subject to the collective and simultaneous action of numerous distributional forces and the dispersion pattern reflects their combined influence. When many small factors act together on the population, a more or less random distribution is to be expected, whereas the domination of a few major factors tends to produce departure from randomness.

**Actions of environmental agencies of transport.** The transporting action of air masses, currents of water, and many kinds of animals produces both random and nonrandom types of dispersion. Air-borne seeds, spores, and minute animals are often scattered in apparently haphazard fashion, but aggregation may result if the wind holds steadily from one direction. Wave action is frequently the cause of large concentrations of seeds and organisms along the drift line of lake shores. The habits of fruit-eating birds give rise to the clusters of seedling junipers and cherries found beneath such perching sites as trees and fenceposts, as well as to the occurrence of isolated individuals far from the original source. Among plants, it seems to be a gen-

eral principle that aggregation is inversely related to the capacity of the species for seed dispersal.

**Physical features of the habitat.** Responses of the individuals of the population to variations in the habitat also tend to give rise to local concentrations. Environments are rarely uniform throughout, some portions generally being more suitable for life than others, with the result that population density tends to be correlated directly with the favorability of the habitat. Oriented reactions, either positive or negative, to light intensities, moisture gradients, or to sources of food or shelter, often bring numbers of individuals into a restricted area. In these cases, aggregation results from a species-characteristic response to the environment and need not involve any social reactions to other members of the population (see ENVIRONMENT).

**Influence of temporal changes.** In most species of animal, daily and seasonal changes in weather evoke movements which modify existing patterns of dispersion. Many of these are associated with the disbanding of groups as well as with their formation. Certain birds, bats, and even butterflies, for example, form roosting assemblages at one time of day and disperse at another. Some species tend to be uniformly dispersed during the summer, but flock together in winter. Hence temporal variation in the habitat may often be as effective in determining distribution patterns as spatial variation.

**Behavior patterns in reproduction.** Factors related to reproductive habits likewise influence the dispersion patterns of both plant and animal populations. Many plants reproduce vegetatively, new individuals arising from parent rootstocks and producing distinct clusters; others spread by means of rhizomes and runners and may thereby achieve a somewhat more random distribution. Among animals, congregations for mating purposes are common, as in frogs and toads and the breeding swarms of many insects. In contrast, the breeding territories of various fishes and birds exhibit a comparatively regular dispersion.

**Intensity of competition.** Competition for light, water, food and other resources of the environment tends to produce uniform patterns of distribution. The rather regular spacing of trees in many forests is commonly attributed largely to competition for sunlight, that of desert plants for soil moisture. Thus a uniform dispersion helps to reduce the intensity of competition, while aggregation increases it.

**Social factors.** Among many animals the most powerful forces determining the dispersion pattern are social ones. The social habit leads to the formation of groups or societies (see SOCIAL ANIMALS). Plant ecologists use the term society for various types of minor communities composed of several to many species, but when the word is applied to animals it is best confined to aggregations of individuals of the same species which cooperate in their life activities. Animal societies or social groups range in size from a pair to large bands, herds, or colonies. They can be classified functionally as mating societies (which in turn are monogamous or

polygamous, depending on the habits of the species), family societies (one or both parents with their young), feeding societies (such as various flocks of birds or schools of fishes), and as migratory societies, defense societies, and other types. Sociality confers many advantages, including greater efficiency in securing food, conservation of body heat during cold weather, more thorough conditioning of the environment to increase its habitability, increased facilitation of mating, improved detection of, and defense against, predators, decreased mortality of the young and a greater life expectancy, and the possibility of division of labor and specialization of activities. Disadvantages include increased competition, more rapid depletion of resources, greater attraction of enemies, and more rapid spread of parasites and disease. Despite these disadvantages, the development and persistence of social groups in a wide variety of animal species is ample evidence of its over-all survival value. Some of the advantages of the society are also shared by aggregations that have no social basis.

**Optimal population density.** The degree of aggregation which promotes optimum population growth and survival, however, varies according to the species and the circumstances. Groups of organisms often flourish best if neither too few nor too many individuals are present; they have an optimal population density at some intermediate level. The condition of too few individuals, known as undercrowding, may prevent sufficient breeding contacts for a normal rate of reproduction. On the other hand, overcrowding, or too high a density, may result in severe competition and excessive interaction that will reduce fecundity and lower the growth rate of individuals. The concept of an intermediate optimal population density is sometimes known as Allee's principle. [F.C.E.]

**Bibliography:** W. C. Allee, *Animal Aggregations: a Study in General Sociology*, 1931; P. Greig-Smith, *Quantitative Plant Ecology*, 1957.

## Population dynamics

The aggregate of processes that determine the size and composition of any population. In this context, a population is considered to consist of organisms of a single species. The group is characterized by definite time rates of birth and death and often by a definite composition with respect to the ratio between the sexes and between the numbers of individuals belonging to different age classes. An aggregation of individuals brought together fortuitously may or may not constitute a population in this sense.

**Population size and density.** Population size is normally measured in terms of numbers of individuals, while productivity is often expressed as the number of new individuals produced per unit time. There are exceptions such as stands of timber or populations of commercially valuable fish. In these instances, productivity and population size are appropriately measured in terms of mass or volume rather than numbers. The study of dynamics in such



populations merely requires consideration of individual growth rates in addition to numbers and ages, and we will here limit our discussion to populations measured by enumeration.

In practice, it is difficult to define the limits of a population, and enumeration normally measures some type of population density. Crude density is the number of individuals per unit of selected space or volume. Examples of this are the number of deer in a county or other areal unit, or the mean number of fish per acre of water surface or per cubic meter of water. This measure suffers from the fact that the units of area and volume are heterogeneous and the population does not utilize all of the available space. Ecological or economic density refers to the mean number of individuals per unit of space actually utilized. Sometimes it is easiest to enumerate a population under conditions of maximum density as when fur seals or colonial birds are on their breeding grounds, when deer are in winter yards, or when snakes are congregated in dens. Often, the only practicable measures of natural populations involve relative density and are designed to show whether a population is increasing or decreasing without determining its actual size. Thus, the number of birds seen per man-hour of walking and the number of squirrels treed per dog-hour have been used to compare population densities in different years and places. Formidable statistical and practical problems are involved in the mensuration of natural populations.

**Population growth.** A population can gain in numbers only by birth and immigration, and it can decrease only by death and emigration. In considering theoretical population dynamics, we customarily ignore migratory movements and concentrate our attention on the birth and death processes. Individuals of every species have the potentiality for producing more offspring than are required to replace the parents. Without this potential, the species could not meet emergencies and would necessarily become extinct. Some organisms reproduce once per lifetime, others many times. Some produce tremendous numbers of gametes, others few. Also, the age at which reproductive maturity occurs varies tremendously—from a few minutes in bacteria to more than a century in the giant *Sequoia* tree. These life history features determine the potential growth rate of the population or the biotic potential of the species.

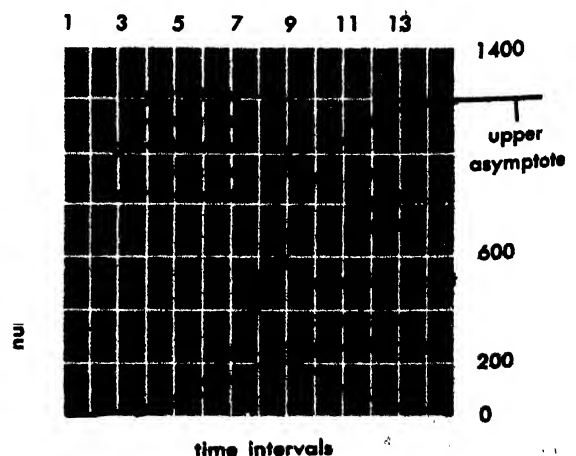
If these life history features remained the same in successive generations, the population would ultimately grow in accordance with the equation  $N_t = Ae^{rt}$  for which the growth rate at time  $t$  is  $dN/dt = rN_t$ . In this formula  $N_t$  is the population size at time  $t$ ,  $A$  is a constant,  $e$  is the base of the natural logarithmic system, and  $r$  is a measure of population growth. The value of  $r$ , which is determined by natural history features, is commonly referred to as the intrinsic rate of natural increase, and it has been proposed to define biotic potential as the normal maximum value of  $r$  for a given population.

This exponential form of potential population growth implies exceedingly rapid expansion. A single individual of an annual plant in which each individual produces only two viable seeds would leave 1,000,000 descendants in 20 years if all seeds survived, and, in fact, every species is theoretically capable of overflowing the earth. In actuality, shifts in natality (birth rates) and mortality inhibit unlimited exponential growth.

**Population control or regulation.** Since potential population growth is exponential, while real population size is limited, much interest and controversy has centered about the form of actual population growth. Commonly, when the size of a growing population is plotted against time, the result is a broad S-shaped (sigmoid) form of growth curve which often appears to be symmetrical about a central point of inflection, where the rate of growth shifts from increase to decrease.

Attempts to give a generalized mathematical formulation of population growth have led to consideration of equations of the form  $dN/dt = rNf(N)$ . This equation differs from that describing exponential growth in that the factor  $f(N)$  serves as a governor or damping factor which takes the value zero when  $N$  is very large, indicating that there is some finite upper limit to the size of a population that is capable of further growth.

Various proposals have been offered concerning the form of this governing factor. If  $f(N)$  is assumed to be merely a decreasing function of time the result is one form of the discredited doctrine of racial senescence, the supposition that populations age and die as do individual organisms. Numerous workers have supposed that  $f(N)$  is a random variable, sometimes positive and sometimes negative, so that populations normally fluctuate about a steady-state, or equilibrium, size. Modern probability theory shows this position to be untenable in so simple a form. By this doctrine, random extinction would be the eventual fate of all populations, but contrary to experience, small populations of obscure forms often should be observed to grow to



Population dynamics. Three types of growth curves using the same data.

tremendous size, at least approaching the condition of overflowing the earth.

In a somewhat more realistic approach, the governing factor is regarded as a decreasing function of population density so that the population inhibits its own growth beyond a certain size. It is obvious, however, that maximum population size must also be affected by the quality of the habitat. Populations of the same species are commonly more dense in some regions than in others owing to differences in the availability of essential resources such as food and nesting sites. Therefore, it is usual to speak of a carrying capacity for any given habitat and to define this as the maximum steady-state population of a given species that can be supported there.

*Carrying capacity.* No general agreement has been reached on a precise definition of carrying capacity, but it seems essential to recognize that this hypothetical upper limit may sometimes be exceeded temporarily. For example, populations as dense as 17 individuals per square yard have been observed in house mice, and the hordes of locusts, chinch bugs, lemmings, and other animals that occur in outbreak years far exceed the capacity of the occupied land to provide food and shelter.

*Environmental resistance.* In practice, students of natural populations often think of population density not in absolute terms but as a density, relative to some standard, such as the maximum that could be supported. Thus, the factor,  $f(N)$ , governing population growth is most realistically regarded as a function of the unfulfilled possibilities for growth. The term environmental resistance has been employed to express this same concept that the resistance to further growth increases as the environment approaches saturation.

It is known from observations in the field and from laboratory experiments that crowding does promote increased death rates and inhibit reproduction. Pathogenic organisms can spread rapidly through populations where there is close contact between individuals, metabolic wastes may accumulate to toxic levels, malnutrition or other deficiency conditions may weaken the individuals, and in very crowded populations there may be interference with feeding and mating. Also, various symptoms of physiological stress apparently result from crowding. There are, therefore, sound reasons for considering increased population density to operate in limiting and finally inhibiting population growth; that is to say, the growing population exhibits negative feedback. Theoretically also, any cause of mortality or sterility that is always independent of population size or density, in the mathematical sense, could not prevent growing populations from occasionally overflowing the earth.

*Density factors.* Numerous students have tried to classify environmental influences into density-independent factors which are theoretically incapable of regulating population size and density-dependent factors which can exert a governing effect. Much controversy has surrounded these concepts and the

varying definitions given to the various classes of factors. Thus, it has been claimed that a density-dependent factor must be what others have called density-responsive; that is, the intensity of the factor must be altered by changing population density. Since population density cannot ordinarily be considered to alter weather or climate, it has been contended that meteorological conditions are density-independent and cannot regulate population size. Others have recognized that weather does sometimes operate in a density-dependent manner, or they have maintained that the true density-dependent regulating factor in such cases is competition for shelter between the individuals which forces the losers to be exposed to unfavorable weather conditions. Another prominent school of thought is utterly opposed to this general approach and maintains that populations seldom attain a level where density effects are important but are normally held at lower levels by environmental inadequacies and mortality factors such as extremes of weather.

*Underpopulation effects.* At the other end of the scale of size there is also a great deal of evidence in many species for underpopulation effects on population growth. In sparse populations, females may have difficulty finding mates. A small population lacks the adaptability to changed conditions that is provided by a large stock of genetic variability, and this defect may be aggravated by inbreeding. Also, populations often condition their surroundings and modify the impact of environmental factors. The forest literally protects the trees from wind damage excessive insulation, and evaporation. The advantages of maintaining the population above some minimum level are obvious in gregarious animals such as bird flocks and herds of ungulates, and even more so in the social insects where ants, bees, and termites, for example, exercise considerable control over the climate inside the colonial structure. In addition, many cases are known where, often for obscure reasons, populations seem unable to resist extinction if the numbers fall below some minimum level.

From these observations it follows that any generalized concept of the growth governor  $f(N)$ , must provide for this factor to be small in very small populations, to rise to a maximum at some optimum population size, and to decline to zero before the population overflows the earth. In other words, the growing population may exhibit positive feedback up to a certain size range and negative feedback at higher levels. To date, however, very little use has been made of highly generalized governing factors.

*Actual population growth.* The logistic function has been the equation most employed for representing actual population growth. This equation in its differential form is

$$\frac{dN}{dt} = rN \left(1 - \frac{K}{N}\right)$$

where  $K$  represents the upper asymptote or maximum size attainable by the population in question. Here the intensity of the governing factor decreases

linearly with population size so no account is taken of underpopulation effects. The integrated logistic curve, however, is sigmoid-shaped and symmetrical about its central point of inflection. It often gives a very good representation of the course of population growth. It occupies a prominent, though controversial, position in modern theories of population dynamics.

**Optimum yield.** An important consequence of the sigmoid form of population growth is the fact that populations of intermediate size are capable of more rapid growth and greater productivity than are either very large or very small populations. If growth were strictly logistic, the most rapid growth would occur at a population size of  $K/2$  and the growth rate at this point would be  $rK/4$ .

When man begins to exploit a large population, as in commercial fishing, the effects of his catch will be to reduce population size. If exploitation is not too intense, the smaller population will lie on a steeper portion of the sigmoid growth curve and will therefore be more productive than the larger population. The population is said to compensate for the increased mortality. In theory, productivity will increase with rate of exploitation to the point where population size reaches the inflection in the growth curve. Hence the maximum possible sustained harvest would be obtained by reducing the population to the inflection point and harvesting at a rate just sufficient to maintain this size. There are many practical difficulties in all actual attempts to determine the optimum rate of harvest and the general problem has become widely known as the optimum-yield problem. It is noteworthy that if the population is "overfished" so that its size passes below the inflection point, productivity will decline with each further decrease in size. Then each increase in the effort to harvest a crop will have the effect of reducing the long-term yield. There is reason to believe that many commercial fisheries are reducing their total catch by fishing too intensively.

The same principles apply to attempts to control noxious species. Rodent populations, for example, compensate for mortality and it is possible to harvest a large annual crop of rats without actually reducing the population. Programs of killing are often discontinued before the population passes below the inflection point where control would become progressively easier. Consequently, the most effective way of dealing with noxious forms is often to reduce the carrying capacity of the environment; programs for improving garbage disposal and for ratproofing buildings will often be much more effective than programs of killing.

**Fluctuations.** Populations of many species fluctuate in size from year to year, and a very large literature exists on this subject. Plagues of rodents and locusts are recorded in the Old Testament and similarly ancient sources, and the migrations of the lemmings and the eruptions, outbreaks, or gradations of various insect populations have often

attracted popular attention. There is a tendency for eruptions to be most conspicuous in high latitudes and other regions where the biota is composed of relatively few species of plants and animals. Outbreak years typically follow periods of build-up during which the population nearly realizes its potential of exponential growth. Eventually, population size exceeds the capacity of the environment to sustain it and the population "crashes," often dropping abruptly to a very low level.

**Cycles.** The most discussed of the fluctuating populations have been those of certain gallinaceous birds, rodents, rabbits, and fur-bearing mammals of northern regions. The records of the Hudson's Bay Company, for example, provide figures for a long series of annual catches, and many students of populations have considered that the rhythms, or cycles, in such records indicate a regular periodicity in the rise and fall of population size. Although cycles of various length have been postulated, most competent opinion in recent years has considered that there are two predominant cycle lengths: a short cycle of approximately 3 or 4 years and a longer cycle often referred to as the 10-year cycle.

Numerous explanations have been advanced. One of the most popular has been the belief that the populations follow some extraterrestrial rhythm, especially the "sun-spot cycle." Such hypotheses suffer from numerous observations indicating that populations in different regions may be out of phase with each other. Others have based explanations on population dynamics, claiming, in effect, that the cyclic species are deficient in feedback mechanisms so that exponential growth is not inhibited until disastrously high densities are attained. Still others have attributed the cycles to interactions between two species: herbivores and their food plants, or predators and their prey. The predator is visualized as growing until it exhausts its food supply and then undergoing violent decline until the prey population has time to recover. These hypotheses are not entirely satisfying because it is difficult to see why many species with diverse life histories should adhere to two basic cycle lengths. The Canadian lynx and the chinch bug, for example, are both considered to exhibit 10-year cycles.

**Other factors.** It has also been noted that random variables such as the sizes of the numbers turning up on a roulette wheel will, when plotted on graph paper, give an appearance of regularity and show a series of peaks occurring at a mean interval of 3-4 numbers. It has been postulated that the great variety of haphazard factors affecting population size constitute a causal system comparable in complexity to that governing the roulette wheel and that the appearance of peak population years may therefore be considered to be governed by a random variable.

Whatever may be the causes of population fluctuations, they are often of great practical importance. Much remains to be learned about possibilities for predicting peak years in crop damage and

in the harvest of food, game, and fur-bearing animals, or for minimizing the expectation of financial loss resulting from these fluctuations.

**Age structure.** Not only the size but also the composition of a population is governed by the age schedules of natality and mortality. If the life history features and the death rates for individuals of each age remain constant, a population will eventually attain a stable age distribution such that the individuals of any particular age constitute a fixed proportion of the total. In human populations, about one-half of the individuals typically fall in the age range of 15-50 years, but the ratio of older individuals to very young differs greatly from one nation to another. Consequently, diseases of old age attain greatest importance in populations where life expectancy is greatest. Problems relating to age structure are also of great significance in nonhuman populations. It is apparent that a predator can benefit man by selectively killing superannuated game animals, thus increasing productivity, or that such a predator can work against man's efforts to control noxious species even while killing many of the undesirable forms. See *ECOLOGV*. [L.C.CO.]

**Bibliography:** Population studies: animal ecology and demography, *Cold Spring Harbor Symposia Quant. Biol.*, vol. 22, 1957.

## Population genetics

The study of both experimental and theoretical consequences of Mendelian heredity on the population level, in contradistinction to classical genetics, which deals with the offspring of specified parents on the familial level. The genetics of populations studies the frequencies of genes, genotypes, and phenotypes, and the mating systems. It also studies the forces that may alter the genetic composition of a population in time, such as recurrent mutation, migration, and intermixture between groups, selection resulting from genotypic differential fertility, and the random changes incurred by the sampling process in reproduction from generation to generation. This type of study enables one to gain an understanding of the elementary step in biological evolution. The principles of population genetics may be applied to plants, and other animals, as well as men. See *EVOLUTION, ORGANIC*; *MENDELISM*.

**Mendelian population.** A Mendelian population is a group of individuals who interbreed among their members according to a certain system of mating and form more or less a breeding community by themselves. These individuals share a common gene pool, which is the total genic content of the group. A Mendelian population is the unit of study in population genetics. The population may be very large or very small, and is to be distinguished from species or varieties, which may consist of numerous isolated or partially isolated Mendelian populations. Mendelian population is a genetic rather than a taxonomic term. Mendelian populations differ from each other in their genic content or chromo-

somal organization, not necessarily in their taxonomic features. The term deme, originally defined as an assemblage of taxonomically closely related individuals, has been used as a synonym for Mendelian population. Camodeme, a deme forming a more or less isolated local intrabreeding community, would be a better substitute.

**Mutation pressure.** Gene mutation arises from time to time in nature. The causes for mutation are not fully known, and thus it can be said that mutations arise "spontaneously." The effect of a new mutant gene is unpredictable and the gene is therefore said to mutate "at random." One property of mutation has been established: it is recurrent. Each type of gene mutates at a certain rate per generation. The rate is usually low—about 1 mutant in  $10^5$ - $10^8$  genes of a given sort, varying from locus to locus on the chromosomes, even under uniform conditions. Ionizing radiation, certain chemicals, heat, and some other agents increase the rate of mutation. See *MUTATION*.

Let  $\mu$  be the rate of mutation from an allele  $A$  to another form  $a$  per generation. If a fraction  $p$  of the genes of a population is  $A$  in one generation, then in the next generation the frequency of  $A$  will be diminished by the amount  $p\mu$ , so that the new frequency of  $A$  will be  $p(1 - \mu)$ . The amount of change,  $p\mu$ , is said to be due to the mutation pressure. If this pressure is unopposed generation after generation, the gene  $A$  will gradually disappear from the population, as  $p_n = p_0(1 - \mu)^n = p_0e^{-n\mu}$  where  $p_0$  is the initial gene frequency and  $p_n$  is the frequency after  $n$  generations. Therefore, for all existing genes, there must be some kind of compensating mechanism which supports its continuing presence in nature. One important problem in population genetics is the mechanism of maintenance of a gene in a population or of its change in frequency from generation to generation.

If, in addition to the mutation from  $A$  to  $a$ , there is reverse mutation from  $a$  to  $A$  at the rate  $\nu$  per generation, then the net amount of change in the frequency of  $a$  is  $\Delta q = p\mu - q\nu$ . At the time when these opposing changes cancel each other, there will be no change in gene frequency despite the recurrent mutations. This state of affairs is said to be in equilibrium and is obtained when  $\Delta q = 0$ ; that is,  $\hat{p} = \nu/(\mu + \nu)$  and  $\hat{q} = \mu/(\mu + \nu)$ , where  $q$  is a frequency of  $a$ ,  $\hat{q}$  is the equilibrium point for  $a$ , and  $\hat{p}$  is the equilibrium point for  $A$ . The equilibrium gene frequencies are determined by the opposing rates of mutation only and are independent of the initial frequencies of the genes in the population. The amount of change in gene frequency per generation is larger when the current  $q$  is far away from the equilibrium  $\hat{q}$  than when  $q$  is close to  $\hat{q}$ . Substitution gives  $\Delta q = -(\mu + \nu)(q - \hat{q})$ , indicating that the amount of change per generation is proportional to the deviation  $(q - \hat{q})$ . It also shows that if  $q > \hat{q}$ ,  $q$  decreases, and if  $q < \hat{q}$ ,  $q$  increases, or that  $q$  will approach  $\hat{q}$  from either side. Such an equilibrium is said to

be stable. The changes in  $q$  described above are independent of the mating system practiced in the population.

In nature, and under artificial conditions, the mutation rates may not remain constant in all generations but may fluctuate within a certain range from time to time. In such cases, instead of a single fixed equilibrium point  $\bar{q}$ , there will be an equilibrium distribution of  $q$  within a certain range, and the apparent change in gene frequency from one generation to the next may be purely a stochastic phenomenon without necessary long-term significance. The same remark applies to all equilibria to be established in subsequent paragraphs. See STOCHASTIC PROCESS.

**Migration and intermixture.** If a fraction  $m$  of a population with a gene frequency  $q$  consists of immigrants from outside and the immigrant group has a gene frequency  $\bar{q}$ , then the new gene frequency of the population will be  $q_1 = (1 - m)q + m\bar{q} = q - m(q - \bar{q})$ . The amount of change in gene frequency in one generation is thus  $\Delta q = q_1 - q = -m(q - \bar{q})$ , showing that the change is proportional to the deviation  $(q - \bar{q})$ . This expression for  $\Delta q$  is of the same form as that for mutation. If the immigrants have the same gene frequency as that in the population, there will be no change in gene frequency in spite of the migrations. The continued intermixture of neighboring populations will eventually make them homogeneous in terms of gene frequencies. Thus, if a large population is divided into a number of partially isolated subpopulations, migrations between the groups will eventually make all subpopulations have the same gene frequency  $\bar{q}$ , which then denotes the average for the entire population in the absence of other disturbing factors. If the local populations are differentiated genetically, there must be some mechanism (for example, local selection) to counteract the pooling effect of migrations so that an equilibrium condition may be reached. The change in gene frequency due to migration is independent of the mating system practiced in the population.

**Mating systems.** In a gene pool with respect to one locus, if a proportion  $p$  of the genes is  $A$  and a proportion  $q$  of the genes is  $a$ , the genotypic proportions in the population are still unknown until the mating pattern is specified. The mating pattern is a system by which the genes are associated into pairs to form the diploid genotypes. The mating systems vary widely in nature for different organisms and populations. Thus, wheat may have 1% cross pollination and 99% self-fertilization, whereas maize practices just the reverse. One of the simplest and most extensively studied systems is random mating, also known as panmixis.

**Panmixis.** Random mating between individuals is equivalent to a random union of gametes. Thus, if the  $(pA, qa)$  gametes of one sex unite at random with the  $(pA, qa)$  gametes of the opposite sex, the resulting genotypic array will be  $p^2AA$ ,  $2pqAa$ ,  $q^2aa$ . These genotypic proportions will be realized

only in very large populations. See HARDY-WEINBERG FORMULA; HUMAN GENETICS.

**Inbreeding.** Inbreeding refers to mating between genetically related individuals; the frequency with which two  $A$  gametes unite will be greater than  $p^2$ ; and a similar situation is true for gene  $a$ . Consequently, inbreeding leads to an increase of homozygosis at the expense of heterozygosis. Let  $H$  be the heterozygosis proportion in a population,  $H'$  that in the preceding generation,  $H''$  that two generations ago, and so on. On continued systematic inbreeding, the manner in which the value of  $H$  decreases is shown in Table 1.

Continued close inbreeding, such as those degrees indicated in Table 1 and many others, eventually leads to complete homozygosis. The population will then consist of  $pAA$  and  $qaa$ ; that is, an  $A$  gamete will always unite with another  $A$  gamete, and an  $a$  gamete with another  $a$ . Inbreeding between remote relatives does not necessarily lead to complete homozygosis but only decreases the heterozygosis below the random mating level to a certain extent.

The inbreeding coefficient is an index intended to measure the amount or degree of inbreeding that has been accomplished in a population. Various indices may be constructed. One that has been proved highly useful in both theoretical investigation and practical breeding work is the inbreeding coefficient  $F$  defined as the correlation coefficient between the uniting gametes. The value of  $F$  ranges from 0 for random mating to 1 for inbreeding in a homozygous population, as shown in Table 2.

In an inbred population where the correlation between the uniting gametes is  $F$ , the genotypic array in the population will be

$$\begin{aligned} AA &: p^2 + Fpq = (1 - F)p^2 + Fp \\ Aa &: 2pq - 2Fpq = 2(1 - F)pq \\ aa &: q^2 + Fpq = (1 - F)q^2 + Fq \end{aligned}$$

The last set of expressions shows that the population may be mathematically considered as having two separate components,  $(1 - F)$  panmictic and  $F$  fixed. If the mating system is such that  $F$  remains constant (instead of increasing) from generation to generation, the population will reach an equilibrium state with the genotypic array shown above.

Table 1. Decrease in heterozygosis with systematic inbreeding

Inbreeding system	Manner of heterozygosis ( $H$ ) decrease	Limiting situation
Self-fertilization	$H = \frac{1}{2}H'$	$H = 0.500H'$
Same purebred sire $\times$ successive daughters	$H = \frac{1}{2}H'$	$H = 0.500H'$
Brother $\times$ sister	$H = \frac{1}{2}H' + \frac{1}{4}H''$	$H = 0.809H'$
Younger parent $\times$ offspring	$H = \frac{1}{2}H' + \frac{1}{4}H''$	$H = 0.809H'$
Half brother $\times$ half sisters	$H = \frac{3}{4}H' + \frac{1}{4}H''$	$H = 0.890H'$
Half brother $\times$ full sisters	$H = \frac{1}{2}H' + \frac{1}{4}H'' + \frac{1}{4}H'''$	$H = 0.970H'$
Double first cousins	$H = \frac{1}{4}H + \frac{1}{4}H'' + \frac{1}{4}H'''$	$H = 0.980H'$

Table 2. Range in value of  $F$ , the inbreeding coefficient

Random mating			
A		a	
A	$p^2$	$pq$	$p$
a	$pq$	$q^2$	$q$
	$p$	$q$	1
Correlation = 0			
Inbred population			
A		a	
	$p^2 + Fpq$	$pq - Fpq$	
	$pq - Fpq$	$q^2 + Fpq$	
	$p$	$q$	1
Correlation = $F$			
Inbreeding in homozygous population			
A		a	
A	$p$	0	$p$
a	0	$q$	$q$
	$p$	$q$	1
Correlation = 1			

The correlation between uniting gametes is due to the correlation between mating individuals. In an equilibrium population, if  $M$  denotes the correlation between mates, then  $M = 2F/(1 + F)$  or  $F = M/(2 - M)$ .

**Genotypic selective values.** Within a large population not all individuals produce the same number of offspring. In the situation to be considered, the average numbers of living offspring born to each of the genotypes in the population are studied, while ignoring the random fluctuation in the number of offspring from family to family. Furthermore, it is assumed that the population is so large that only the relative frequencies of the various genotypes and genes in the population are of interest. Suppose that the average number of offspring for each genotype is as follows:  $AA$ : 2.00;  $Aa$ : 2.50;  $aa$ : 1.50. Given these differential rates of reproduction, the new gene frequency of the next generation may be calculated. Inasmuch as it is only their relative magnitude that matters, these reproductive rates may be simplified into the ratio  $W_{11}:W_{12}:W_{22} = 1:1.25:0.75$  or alternatively into 0.80:1:0.60. In order to standardize the description, it is convenient to take one of the three reproductive values as unity. In the previous example, depending upon whether the reproductive value of  $AA$  or  $Aa$  is taken as unity (the standard), that of  $aa$  is 0.75 or  $0.60 = 1 - 0.25$  or  $1 - 0.40 = 1 - s$  in general. The value of  $s$  is known as the selection coefficient against the genotype  $aa$ . When a selection coefficient is used, it should always be stated which genotype has been employed as the standard.

**Natural selection.** The doctrine of the survival of the fittest needs clarification from the genetic view-

point. The relative genotypic reproductive values ( $W_{11}, W_{12}, W_{22}$  of the preceding paragraph) simply give an ex post facto description, by which the genetic composition of the offspring generation may be related to that of the parent generation. These  $W$  values include all causes for differential reproduction such as fecundity, fertility, sexual maturity and capacity, survival and viability, length of reproductive life, and many others, depending on the details of the life cycle of the organism. The  $W$  value, sometimes briefly referred to as relative "fitness," is not necessarily correlated with any observable morphological characteristics, no matter how desirable they may seem to man. From the genetic viewpoint, only those who reproduce count. Thus, natural selection has no particular purpose except to perpetuate those who are fit to reproduce under the given conditions. Only when a characteristic lowers the organism's reproductive capacity does it have a genetic effect on the subsequent generations.

**Selection pressure and equilibrium.** The effect of selection may be described in terms of changes in gene frequency. In a random mating population with respect to one gene locus, the situation is as shown in Table 3.

The population after selection is the parental population of the next generation through random mating. The value  $\bar{W}$  is the total of the selected parental population, but may also be regarded as the average fitness of the original unselected population. The new frequency of gene  $a$  among the selected is  $q' = (pqW_{12} + q^2W_{22})/\bar{W}$  and therefore the amount of change per generation is  $\Delta q = q' - q$ , or, more explicitly,

$$\Delta q = \frac{pq}{2\bar{W}} \cdot \frac{d\bar{W}}{dq}$$

This represents the effect of selection pressure on gene frequency. When  $p$  or  $q$  is zero, there is no change in gene frequency; there can be no selection in the absence of alternatives. Therefore, all selection effects involve the factor  $pq$ . Further, when  $p$  or  $q$  is very small, the selection is ineffective whether it is for or against a gene. Besides these terminal conditions, if there exists a  $q$  value such that  $\Delta q = 0$ , it is called the equilibrium value of gene frequency, because a population with that particular gene frequency will remain unchanged in spite of the selection pressure. When such a  $q$

Table 3. Effect of selection in a random mating population

Genotype	Frequency, $f$	Fitness, $W$	Frequency after selection, $fW$
$AA$	$p^2$	$W_{11}$	$p^2W_{11}$
$Aa$	$2pq$	$W_{12}$	$2pqW_{12}$
$aa$	$q^2$	$W_{22}$	$q^2W_{22}$
Total	1.00		$\bar{W}$



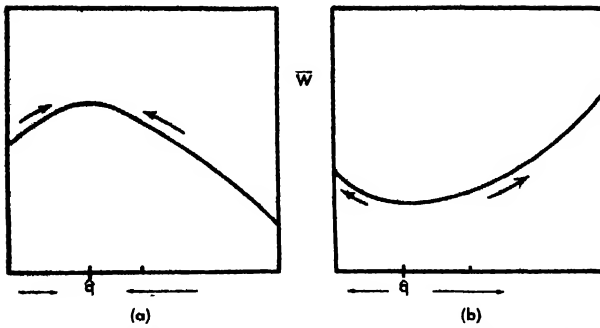


Fig. 1. Diagram of the relationship between the average fitness value  $\bar{W}$  and  $\hat{q}$  the equilibrium point for  $a$ . (a) Stable equilibrium. (b) Unstable equilibrium.

value exists, it must be the following solution of the equation  $d\bar{W}/dq = 0$ :

$$\hat{q} = \frac{W_{11} - W_{12}}{(W_{11} - W_{12}) + (W_{22} - W_{12})}$$

In order that  $\hat{q}$  be a positive fraction, the differences  $W_{11} - W_{12}$  and  $W_{22} - W_{12}$  must be both positive or both negative; that is, the selective value of the heterozygote must be lower or higher than those of both homozygotes. For all other cases, there will be no equilibrium except when  $q = 0$  or 1.

**Stability of an equilibrium.** The value of  $\bar{W} = p^2W_{11} + 2pqW_{12} + q^2W_{22}$  may be plotted against the value of  $q$  or  $p$ . When  $W_{12}$  is greater than  $W_{11}$  and  $W_{22}$ , the  $\bar{W}$  curve has a maximum point (Fig. 1). The  $q$  value corresponding to the maximum value of  $\bar{W}$  is the stable equilibrium point. This means that whether it is smaller or larger than  $\hat{q}$ , the  $q$  value will approach  $\hat{q}$  as selection proceeds from generation to generation. A stable equilibrium of this type leads to balanced genetic polymorphism, that is, to the coexistence of alleles in a population. Conversely, if  $W_{12}$  is lower than both  $W_{11}$  and  $W_{22}$ , the  $\bar{W}$  curve has a minimum point yielding an unstable equilibrium; the selection pressure will make the  $q$  value move away from the equilibrium value toward either 0 or 1, depending upon which side of the equilibrium the  $q$  happens to be. Consequently, selection against the heterozygote leads to the elimination of one of the alleles. In more complicated situations, there could be more than one stable or unstable equilibrium value in a population, or both.

The genotypic selective values  $W_{11}$ ,  $W_{12}$ ,  $W_{22}$  have been assumed to be fixed for each genotype, but in nature they may vary in various ways. In addition to the omnipresent random fluctuations, the selective values may vary with the gene frequency itself. For instance, a genotype may be favored by selection when it is rare in the population but suffer a disadvantage when it is too common. For such cases, there will be an equilibrium yielding genetic polymorphism. Let  $U_{11}$ , a function of  $q$ , be the varying selective value of genotype  $AA$ , and so on. Then the equilibrium value of gene fre-

quency is given by the appropriate solution of the equation

$$q = \frac{U_{11} - U_{12}}{(U_{11} - U_{12}) + (U_{22} - U_{12})}$$

The study of selection effects may be extended to cases with multiple alleles, sex-linked alleles, autopolyploids, and inbreeding populations.

**Gametic selection.** The effective rate at which the  $A$  and  $a$  gametes function may not be the same; that is, selection may operate in the gametic stage instead of in the diploid genotypic stage. If the selective actions for the genotype  $aa$ , for example, and gamete  $a$  are in opposite directions, an equilibrium may result.

**Balance between selection and mutation.** There are many different types of genotypic selection. Two simple cases will illustrate the principle of balance between selection and mutation pressures.

**Selection against recessives.** Suppose that the selective values of  $AA$ ,  $Aa$ , and  $aa$  are 1, 1, and  $1 - s$ , where  $s$  is a positive fraction known as the selection coefficient. Then the new gene frequency will be  $q' = (q - sq^2)/(1 - sq^2)$  in the next generation, so that the amount of change per generation is  $q' - q = -sq^2(1 - q)/(1 - sq^2)$ . At the same time, if  $\mu$  is the mutation rate from  $A$  to  $a$ , the value of  $q$  will be increased by the amount  $\mu(1 - q)$  per generation. At equilibrium, the forces to increase and to decrease the gene frequency must cancel each other; that is,  $\mu(1 - q) = sq^2(1 - q)/(1 - sq^2)$ . Solving, one obtains  $sq^2 = \mu/(1 + \mu)$ , which closely approximates  $\mu$ . Hence,  $q^2 = \mu/s$  and  $q = \sqrt{\mu/s}$ , which usually is a small quantity. When  $aa$  is lethal or unable to reproduce,  $s = 1$  and  $q = \sqrt{\mu}$ . This explains the persistence of deleterious recessive genes in a population in spite of continuous selection.

**Selection against dominants.** If the selection is against homozygous dominants only, the situation is the same as in the previous instance except for substitution of  $p$  for  $q$  and  $v$  (mutation rate from  $a$  to  $A$ ) for  $\mu$ . To bring out the distinction between selection against dominants and that against recessives, take the extreme case in which  $AA$  is lethal, the selective value of  $Aa$  is  $1 - s$ , and  $aa$  is the norm. The value of  $p$  will then be so low that the usual genotypic proportions  $p^2$ ,  $2pq$ ,  $q^2$  will take the limiting form 0,  $2p$ ,  $1 - 2p$ , as  $q$  is very close to unity. The increase in  $q$  through selection is approximately  $sp$ , whereas the loss through mutation from  $a$  to  $A$  is  $qv = v$ . Hence, at equilibrium,  $p = v/s$ . This value is much lower than  $q = \sqrt{\mu/s}$  for selection against recessives. Thus, selection against dominant alleles is more effective than selection of the same intensity against recessives. Selection against heterozygotes will eventually lead to the same limiting situation. All the equilibrium values supported by mutation pressure are low but stable. Mutations prevent complete extinction of an allele.

**Random drift.** The random drift of gene frequencies in finite populations is often called the Sewall Wright effect because of his analysis of its significance. The gene frequency of any generation is determined by the uniting gametes produced by the parents of the preceding generation. If the number of parents is limited and constitutes a random sample of the entire population, the gene frequency of the next generation will not remain exactly the same as that of the previous generation but will be subject to a random fluctuation on account of the sampling process. In a random mating population of  $N$  individuals, one-half of whom are males and one-half females, and maintaining the same population size, the variance of the gene frequency based on  $2N$  gametes is  $q(1-q)/2N$ . The gene frequency may become a little higher or lower in the following generation. The smaller the population, the greater is the variance. This random process will continue to operate in all generations. In a sufficiently long time, the value of  $q$  will reach either the terminal value 0 or 1. Hence the random drift leads eventually to complete homozygosis for small populations. It can be shown that the limiting rate of reaching the state 0 or 1 is each  $\frac{1}{4N}$  per generation so that the total rate of

"decay" of genetic variability is  $\frac{1}{2N}$  per generation. Naturalists have found numerous small isolated colonies (for example, snails in mountain valleys) with characteristics uncorrelated with the environmental conditions to substantiate the theory of random (nonadaptive) fixation.

The effective size of a population is the actual number of individuals producing offspring and thereby responsible for the genetic constitution of the next generation. The random mating population with one-half males and one-half females, and producing the same number of offspring, is an idealized model. Any deviation from the ideal situation will have a different sampling variance and a different rate of decay. Equating these to the "standard" variance  $q(1-q)/2N$  or the ideal decay rate  $\frac{1}{2N}$ , an equivalent  $N$  is obtained for the ideal population. The latter number is known as the effective size of a population. It is convenient to use in mathematical descriptions of the genetic behavior of a population. Some of the factors that tend to make the effective size smaller than the actual breeding size are given below.

1. Unequal number of males and females. If  $M$  and  $F$  are the respective numbers, the effective size  $N_e$  is not simply  $M + F$  but is defined by  $1/N_e = \frac{1}{4M} + \frac{1}{4F}$  and is equal to  $N_e = 4MF/(M + F)$ . The larger the difference between  $M$  and  $F$ , the smaller the number  $N_e$  as compared with  $M + F$ .

2. Unequal size of families. If the gametes are drawn wholly at random from the parents, the number of gametes  $k$  contributed by a parent will form a Poisson distribution. In such a case, the

effective size is the same as the actual breeding size. However, without perfect random sampling, if the mean number of gametes per parent is  $\bar{k} = 2$ , the effective size is equal to

$$N_e = (4N - 2)/(\sigma_k^2 + 2)$$

where  $\sigma_k^2$  is presumably larger than 2.

3. Inbreeding. If  $F$  is the inbreeding coefficient of a population, then the effective size is  $N_e = N/(1 + F)$ .

4. Periodic change in population size. If  $N_1, N_2, \dots, N_t$  are the respective sizes of the  $t$  generations, the average effective size for the period is approximately equal to the harmonic mean of the  $t$  sizes. The harmonic mean is much closer to the smallest number of a series than to the largest one.

**Gene frequencies.** A stationary distribution of gene frequencies results from two opposing forces: the systematic pressures (mutation, migration, selection) which tend to make the gene frequency attain a certain fixed value, and the random variation due to sampling which tends to make the gene frequency drift away from any fixed value. The result of these opposing tendencies is not a single equilibrium value of gene frequency but a stationary distribution of gene frequencies. This distribution may be viewed in three different ways: as the distribution of  $q$  for a particular locus in a population in a long period of time; as the distribution of the allelic frequencies of all loci subject to the same pressures in one population at any given time; and finally as the distribution of  $q$  of one locus among a large number of populations of the same size and with the same pressures at a given time. See ALLELE.

Under selection pressure and the sampling variation, the distribution function  $\phi(q)$  is  $\phi(q) = C\bar{W}^{2N}/q(1-q)$ , where  $C$  is a constant,  $\bar{W}$  the average fitness of the population, and  $N$  the effective size of the population. The exact form of the distribution depends upon the value of  $\bar{W}$ , which is a function of the selection coefficients as well as the gene frequency.

If there is mutation, or migration pressure, or both, the distribution of gene frequency is simply a  $\beta$  distribution:  $\phi(q) = Cq^{U-1}(1-q)^{V-1}$ , where  $U = 4N\mu$ ,  $V = 4N\nu$  if there is only mutation pressure;  $U = 4Nm\bar{q}$ ,  $V = 4Nm\bar{p}$  if there is only migration pressure; and  $U = 4N(\mu + m\bar{q})$ ,  $V = 4N(\nu + m\bar{p})$  if there are both. The exact form of the  $\beta$  distribution depends upon the values of  $U$  and  $V$ . When they are smaller than unity, the population is considered small; when they are close to unity, the population is intermediate in size; when they are much larger than unity, the population is considered large. When the effects of mutation, migration, and selection are combined, the distribution function is

$$\phi(q) = C\bar{W}^{2N} q^{U-1}(1-q)^{V-1}$$

Figure 2 shows some of the forms of the distribution under various conditions. The joint distribution of more than one locus is naturally very com-

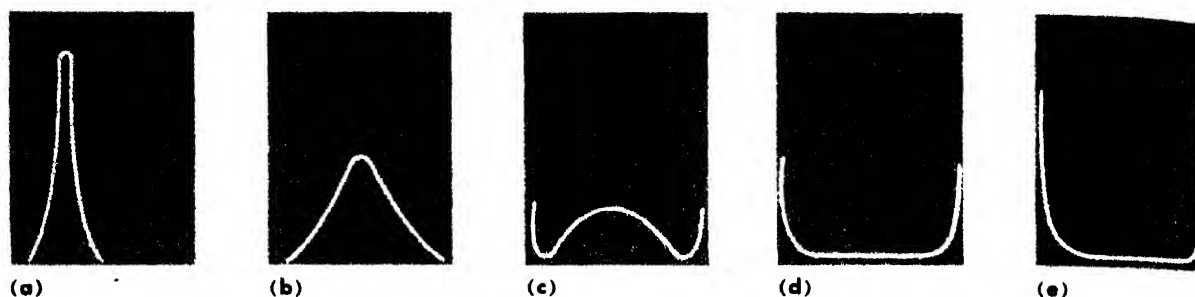


Fig. 2. (a) Large population, the distribution clustered close to the equilibrium value. (b) Intermediate population, a distribution of wider spread. (c) Intermediate small population, a distribution with some terminal fix-

ation. (d) Small population, most genes close to fixation or extinction in the absence of systematic pressures. (e) Small population with strong selection pres-

plicated. Furthermore, if the mutation rates and selection coefficients also vary instead of being constant, the mathematical description of the stochastic process becomes very laborious and the forms shown in Fig. 2 give only a first approximation to the real situation.

The distribution forms of the gene frequencies depend upon the relative magnitudes of the various factors which bring about populational changes. In large random-mating populations all gene frequencies remain close to their stable equilibrium values, which are determined by the counteracting but systematic pressures of mutation, selection, and migration. There will be no further genetic change unless the environmental conditions change so as to define new equilibrium points. Evolution in such large populations is guided essentially by intragroup selection, and progress is very slow.

In small and completely isolated populations most of the gene frequencies are close to 0 or 1 because of the random drift process which dominates the situation. Selection is ineffective. The loci are prevented from being completely fixed only by occasional mutations or immigrants. The ultimate fate of such small homozygous populations is probably extinction because they are nonadaptive and unable to respond to new conditions.

In populations of intermediate size, all factors, both random and systematic, come into play and the population is more responsive to evolutionary change. If a large population is subdivided into many partially isolated groups with migrations between them, there will be some differentiation among the groups, some of it adaptive and some nonadaptive, but there is very little fixation. The selection effect, varying from one locality to another, then operates largely on an intergroup basis which is more efficient than the intragroup selection within one single large population. If the groups are small, some of them will be eliminated by selection while others flourish. This provides the most favorable condition for evolutionary success for the species as a whole. The conclusion is that there is no one all-important factor in evolution. Evolutionary advance depends upon the interplay and balance of all factors. See BIOMETRICS;

EVOLUTION, ORGANIC; GENETICS; MUTATION (ONTOGENY). [C.C.L.]

*Bibliography:* R. A. Fisher, *The Genetical Theory of Natural Selection*, 2d ed., 1959; J. B. S. Haldane, *The Causes of Evolution*, 1932; C. C. Li, *Population Genetics*, 1955; G. Malécot, *Les mathématiques de l'hérédité*, 1948; S. Wright, Evolution in Mendelian populations. *Genetics*, 16:97-159, 1931; S. Wright, The genetical structure of populations, *Ann. Eugenics*, 15(4):323-354, 1951.

## Porcelain

A high-grade ceramic ware characterized by high strength, a white color (under the glaze), very low absorption, good translucency, and a hard glaze. Equivalent terms are European porcelain, hard porcelain, true porcelain, and hard paste porcelain. See GLAZING; POTTERY.

Porcelain is distinguished from other fine ceramic ware, such as china, by the fact that the firing of the unglazed ware (the bisque firing) is done at a lower temperature (1000-1200°C) than the final or glost firing, which may be as high as 1500°C. In other words, the ware reaches its final state of maturity under the glaze. This makes for an exceedingly intimate bond between the glaze and the body, which in turn gives high strength.

The white color is obtained by using very pure raw materials; the low absorption results from the high firing temperature; and the translucency from the chemical composition, the high firing temperature, and the very thin sections in which the ware can be made.

The term porcelain has been applied to the material used to make such things as electrical insulators and bathroom fixtures. Very often these are made in a one-fire process, the glaze being applied to the green or unfired ware; where this is the case and high-grade materials are used in compounding the body, the term porcelain may be correctly applied. However, the pieces have no translucency because of their great thickness. On the other hand, the term porcelain is often applied to quite different ware. For example, zircon porcelain is used to describe a material made largely of zircon ( $ZrO_2 \cdot SiO_2$ ), with small amounts of fluxes to yield

a low absorption; this material might better be called vitrified zircon. See CERAMIC TECHNOLOGY. [M.C.M.]

## Porcupine

Any of several large rodents with some of the hairs modified into pointed quills. The Old World forms are terrestrial, nocturnal animals. The single North American species, *Erethizon dorsatum*, is placed in the family Erethizontidae. It is characterized by its robust body and deliberate movements; it is arboreal and active during the day. The porcupine is found all across the northern United States, most of Alaska and Canada, and southward in the mountains.



The porcupine, *Erethizon dorsatum*; length to 3 ft. (From P. M. Duncan, ed., Cassell's Natural History, Cassell.)

Porcupines do not shoot their quills, but can lash out quickly with the tail and embed many barbed quills in an adversary. They are protected by law in much of the North because of their value as an emergency food for men lost in the woods. In some areas they are looked upon as pests because of the damage they do to trees by girdling them. See RODENTIA. [J.D.B.]

## Porifera

The sponges, a phylum of the animal kingdom which includes about 5000 described species. The body plan of sponges is unique among animals. Currents of water are drawn through small pores or ostia in the sponge body and leave by way of larger openings called oscula. The beating of flagella on collar cells or choanocytes, localized in chambers in the interior of the sponge, maintains the water current. Support for the sponge tissues is provided by calcareous or siliceous spicules, or by organic fibers, or by a combination of organic fibers and siliceous spicules. The skeletons of species with supporting networks of organic fibers have long been used for bathing and cleaning purposes. Because of their primitive organization, sponges are of interest to zoologists as an aid in understanding the origin of multicellular animals. See ANIMAL KINGDOM; PARAZOA.

**Taxonomy.** The Porifera are divided into the Hexactinellida, Calcarea, and Demo-

spongiae on the basis of the skeletal structure. A taxonomic scheme of the Porifera follows. See separate articles on each group.

### Class Hexactinellida

#### Subclass Amphidiscophora

##### Order Amphidiscosa

##### Order Hemidiscosa

#### Subclass Hexasterophora

##### Order Hexactinosa

##### Order Lychiniscosa

##### Order Lyssacinosa

##### Order Reticulosa

### Class Calcarea

#### Subclass Calcinea

##### Order Clathrinida

##### Order Leucettida

##### Order Pharetronida

#### Subclass Calcaronea

##### Order Leucosoleniida

##### Order Sycttida

### Class Demospongiae

#### Subclass Tetractinomorpha

##### Order Homosclerophorida

##### Order Choristida

##### Order Clavaxinellida

#### Subclass Ceractinomorpha

##### Order Dendroceratida

##### Order Dictyoceratida

##### Order Haplosclerida

##### Order Poecilosclerida

##### Order Halichondrida

**General structure and cell types.** The structure and functioning of a sponge is most easily studied in young fresh-water sponges as they develop from dormant bodies called gemmules. Such young sponges possess two discrete canal systems, an inhalant and an exhalant system, which are in communication with each other by way of numerous chambers lined with flagellated collar cells. The outer or dermal epithelium is made up of flattened polygonal cells called pinacocytes and is perforated at intervals by pores or ostia through which water enters the inhalant canal system, usually by way of large subdermal cavities. The exhalant canals are also lined by pinacocytes.

All other cell types in the sponge are located in the inhalant canals. These cells may be divided into two categories, (1) those with large nucleolate nuclei and (2) those with smaller nuclei lacking nucleoli or with very small nucleoli. In the former category are the archaeocytes, ovoid in shape with blunt pseudopodia and hyaline cytoplasm, and many types of wandering amoebocytes which are named according to their functions. They are called chromocytes if they bear either pigment granules or zoochlorellae (unicellular green algae), thesoocytes or spherular cells if they are filled with food reserves, and trophocytes if they are acting as nutritive cells for developing eggs. Mucus-secreting gland cells, presumably derived from amoebocytes, are of common occurrence in sponges. Scleroblasts,

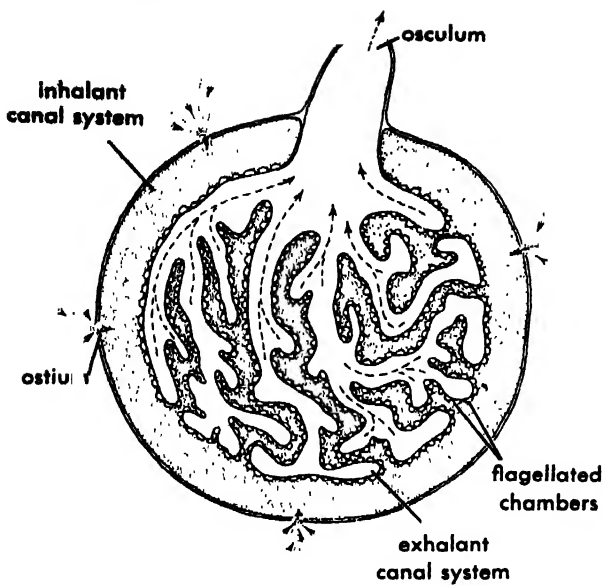


Fig. 1. Diagram of the canal systems of a young fresh-water sponge cultured between cover slip and microscope slide. (After Ankel, 1950)

derived from archaeocytes, secrete spicules, and spongioblasts secrete spongin fibers.

The anucleolate cells with small nuclei include the pinacocytes, mentioned above, collencytes, and choanocytes. In addition to their function as epithelial cells, pinacocytes may be differentiated into spindle-shaped contractile cells, called myocytes, which occur in circlets under the epithelium around ostia and oscula. Collencytes are star-shaped and have long, thin, protoplasmic prolongations which insinuate throughout the inhalant canals. When they are bipolar, collencytes are called desmacytes or fiber cells and are found in abundance in the cortex of some species of sponges. Star-shaped collencytes with a vesicular or vacuolate cytoplasm are called cystencytes. The choanocytes are the most highly differentiated cells of the sponge. Each consists of a small spherical cell body surmounted by a collar formed of many individual contractile cytoplasmic tentacles and provided with a flagellum which extends from the lumen of the collar.

The mesenchymal cells and the skeleton of the Demospongiae and Calcareia are surrounded by a colloidal gel, called mesoglea, secreted by amoebocytes and possibly choanocytes. Elastin fibers which help provide support for the sponge occur in the mesoglea of most species.

**Water-current-system physiology.** The flagellated chambers are hemispherical in shape with a diameter of 40–60  $\mu$  in fresh-water sponges. They are composed of choanocytes firmly held together. The uncoordinated beating of the flagella of the choanocytes creates the flow of water through the sponge. Water enters the flagellated chambers from the inhalant canals by way of openings between choanocytes. Two or three such openings (prosopyles) lead into each chamber. Water leaves each

chamber and enters the exhalant system by way of a single larger pore (apopyle) through the epithelial lining of the exhalant canal. The lumen of an apopyle is ten times greater than that of all the prosopyles leading into the chamber; thus water enters the chambers at a velocity ten times greater than that at which it leaves. The collars of the choanocytes are so oriented that each directs its current toward the apopyle.

A fall in pressure takes place between apopyles and prosopyles and therefore the prosopyles exert a suction effect on the water in the inhalant canals. The effect of this lowered pressure is transmitted to the subdermal cavity, and water is therefore drawn into the ostia forcibly. Food particles which may be at considerable distances from the pores are sucked in along with the water.

The incurrent canals are filled with a network of free cells which tends to decrease the velocity of the water, but the retarding effect of these frictional forces is compensated by the resultant decrease in diameter of the streams of water passing through the inhalant canals. The velocity of current flow through the inhalant canals is twice as great as that through the exhalant canals which have smooth walls but are also greater in diameter. An excess pressure exists in the exhalant canals as a result of the pumping action of the flagellated chambers

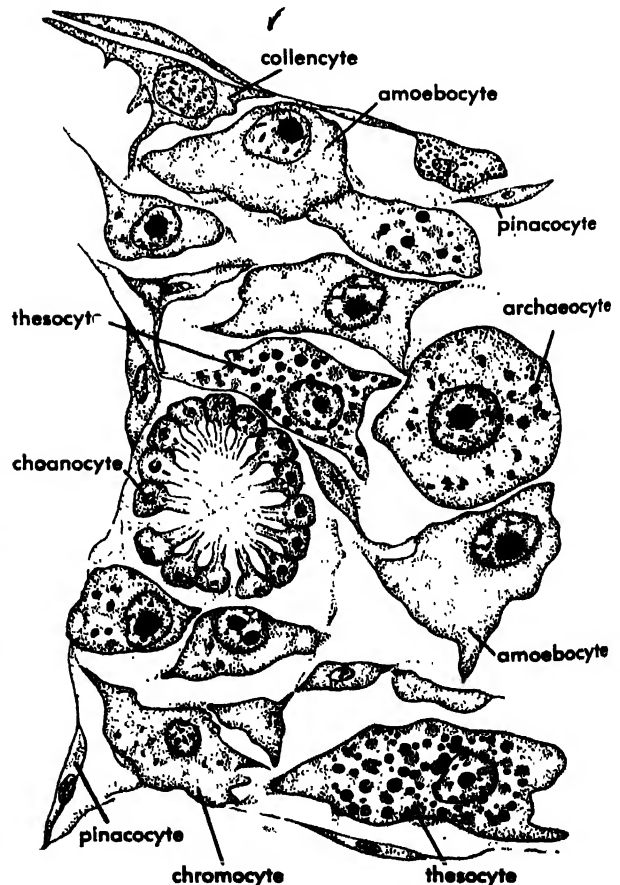


Fig. 2. Types of cells found in a fresh-water sponge as seen in a cross section through the interior of the sponge. (After Meewis, 1936)

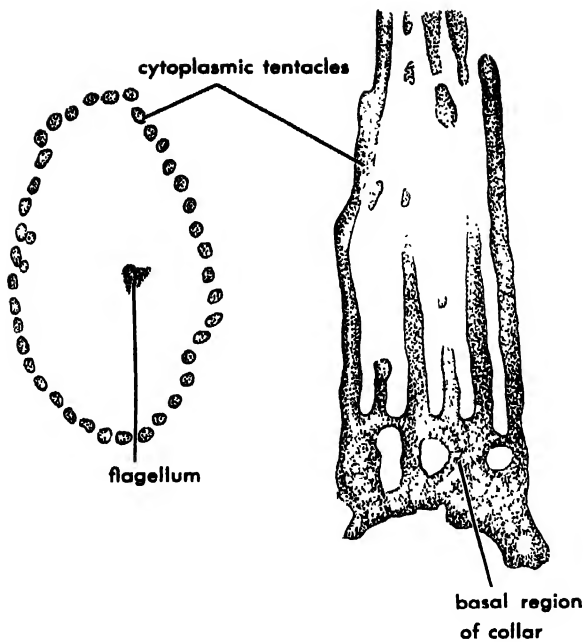


Fig. 3. Transverse (left) and tangential (right) sections through the collar of a choanocyte of a fresh-water sponge. Electron micrographs. (After Rasmont et al., 1957)

and the changing size of the oscular opening. Hence food-free water laden with excretory products is carried far away from the sponge as it leaves the osculum and is unlikely to reenter through the ostia.

**Feeding and digestion.** Feeding activities are best known in fresh-water sponges in which the dermal ostia, the inhalant canals, the flagellated chambers, and the choanocyte collars constitute a set of sieves of decreasing mesh size. The dermal ostia are  $50\ \mu$  in diameter, the prosopyles are  $5\ \mu$  in diameter, and the spaces between the cytoplasmic tentacles of the collars vary from  $0.10$  to  $0.15\ \mu$ . Only the smallest particles can enter the flagellated chambers and come in contact with the collars of the choanocytes. The latter trap the small particles which are passed to the cell body and ingested. Larger particles which have been able to pass through the dermal pores are ingested by the archaeocytes and collencytes which form a reticulum in the inhalant canals or by the cell bases of the choanocytes. Particles too large to enter the ostia can be ingested by the pinacocytes of the dermal epithelium. Observations on the natural food of sponges are few, but scattered reports suggest that unicellular algae, bacteria, and possibly organic detritus constitute the chief items of food.

Direct observations of living fresh-water sponges indicate that the archaeocytes are the chief cells responsible for digestive processes. After food has entered the sponge colony, choanocytes take up and then rapidly lose the particles which they have ingested, and it has been assumed that the particles are transferred to archaeocytes and amoebocytes for digestion. Curiously enough, however, studies

of the activity of digestive enzymes in a marine siliceous sponge indicate much higher proteolytic, lipolytic, and carbohydrate-decomposing activities in choanocytes than in archaeocytes. Perhaps a higher rate of digestive activity accounts for the rapid disappearance of particles in the choanocytes. Observations of the transfer of materials from these cells to archaeocytes may have concerned indigestible products. It is also possible that species differences in digestive functions exist. Following the digestion of the particles in the archaeocytes, these cells migrate to the walls of the exhalant canals or to the dermal epithelium where indigestible material is voided from blisterlike vacuoles.

**Reactions of sponges.** Studies of a shallow-water marine siliceous sponge, *Hymeniacidon heliophila*, have shown that the oscula close upon exposure to air and when the sponge is in quiet sea water. Touching or stroking either oscula or ostia does not induce closure. Ostia remain open in air or when the sponge is in quiet or silt-laden water. The reactions of this sponge are thought to be solely the result of the contractions of myocytes reacting directly to external stimuli. Movements are slow and transmission of stimuli is limited. Reactions of oscula in the fresh-water sponge *Ephydatia fluviatilis* differ somewhat from those of the previous species. Oscula remain open in running or still water. The oscular tubes elongate in still water and flatten out in strong currents. If a needle is rubbed around the edge of the oscular opening the orifice contracts immediately. Weak electrical stimuli applied to the tip of an osculum cause it to close, and a wave of contraction runs down the entire oscular tube. Transmission of stimuli to neighboring regions of the sponge was not observed.

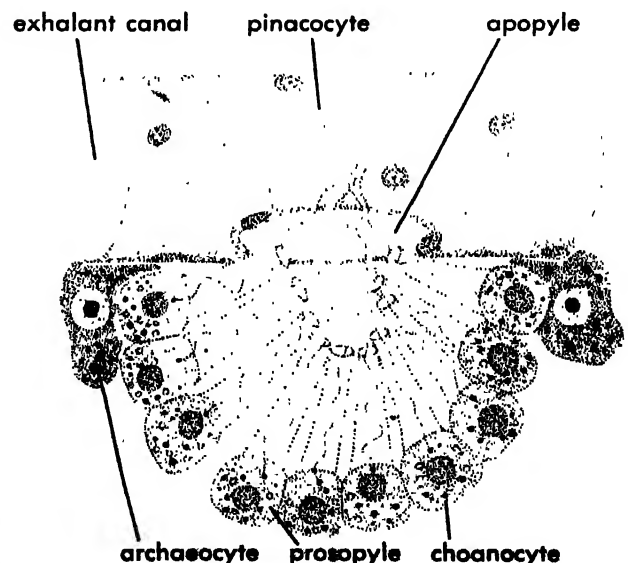


Fig. 4. Diagram of flagellated chamber of a fresh-water sponge. Chamber shown in optical section; exhalant canal in three dimensions. (After Kilian, 1952)



Although neither of the siliceous sponges described above reacted in any way to light, it has been reported that in the calcareous sponge *Leucosolenia* bright light causes oscular constriction and a change in tube shape.

Although there is little evidence from the behavior of sponges that nerve cells are present, cytological studies suggest that such cells do exist. Cells interpreted as being bipolar and multipolar nerve cells have been described from a wide variety of Demospongiae and Calcareae. In fresh-water sponges and in a few marine Demospongiae peculiar cells called lophocytes have been described beneath the dermal membrane. These cells, which bear a process terminating in a tuft or fibrils, may also be nerve cells, although it has been suggested alternatively that they secrete fine fibers forming a layer above them in the case of one species of fresh-water sponge. Unequivocal evidence that nerve cells exist in sponges can come only from physiological studies, and reports based on cytological work alone must be viewed with skepticism at present.

**Skeleton.** Characteristic of sponges is the presence of a skeleton of sclerites or of organic fibers or both. Only a few genera, such as *Halisarca* and *Oscarella*, lack skeletal elements. The skeleton is of primary importance in the classification of sponges; indeed, the three classes of the phylum are separated on the basis of skeletal structure.

The shapes of the sclerites or spicules vary greatly, and an elaborate terminology has been developed for them by taxonomists. An initial sub-

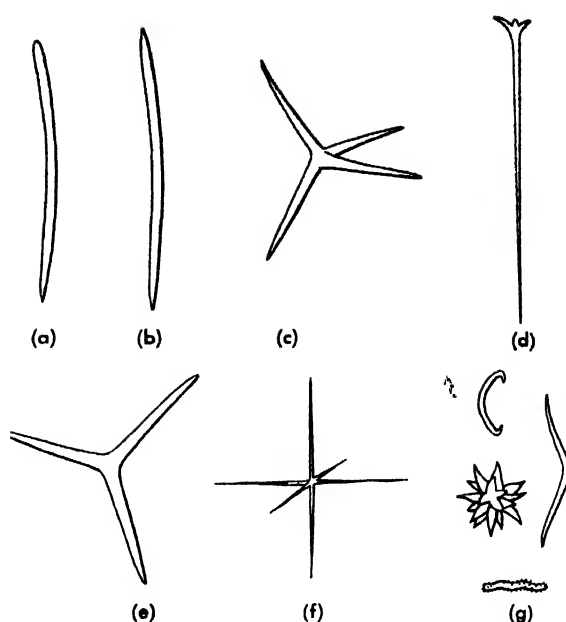


Fig. 6. Spicule types. Monaxons: (a) monactinal and (b) diactinal. Tetraxons: (c) equi-rayed and (d) triaene. (e) Triradial. (f) Triaxon or hexactine. (g) Microscleres.

division into megascleres and microscleres is made on the basis of size. A more detailed nomenclature is based on the number of axes or rays present; an appropriate numerical prefix is added to the endings -axon (referring to the number of axes) or -actine (referring to the number of rays or points). Spicules formed by growth along a single axis are known as monaxons. Such spicules are monactinal (pointed at one end only) if growth occurs in one direction or diactinal (pointed at each end) if growth occurs in both directions. Tetraxons are spicules with four axes or rays, each pointing in a different direction. All four rays may be equal in length, but often one ray is longer than the other three, and the spicules are called triaenes. Triradial (three-rayed) spicules are commonly found in calcareous sponges. Hexactinellid sponges are characterized by the presence of spicules with three axes (triaxons) meeting at right angles. Such spicules have six rays and are also called hexactines.

The fibrous skeletons of bath sponges are composed of spongin, a scleroprotein in which are incorporated halogenated amino acids (monoiodo-, diiodo-, monobromo-, and dibromotyrosine). Spongin is also found in many species of Demospongiae with siliceous spicules where it serves as an interspicular cement or forms fibers in which the spicules are embedded.

**Reproduction.** Both sexual and asexual reproduction occur in sponges. The formation of asexual bodies, gemmules, is characteristic of all fresh-water species.

**Sexual reproduction.** This type of reproduction occurs in most sponges, but it has been reported in only a few species of the order Choristida of the class Demospongiae. The germ cells may arise

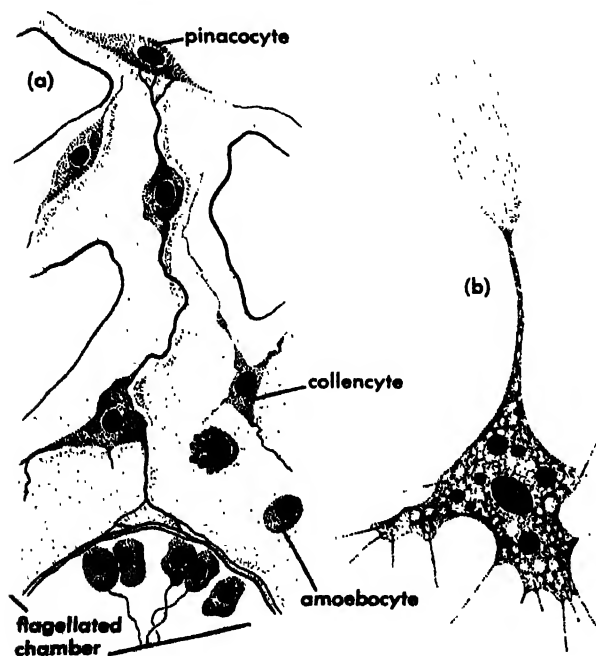


Fig. 5. Supposed nerve cells in sponges. (a) Two-cell arc joining pinacocyte and flagellated chamber, calcareous sponge. (b) Sensory cell near surface connected to "neuron" in mesenchyme, calcareous sponge (after Pavans de Ceccatty, 1955).

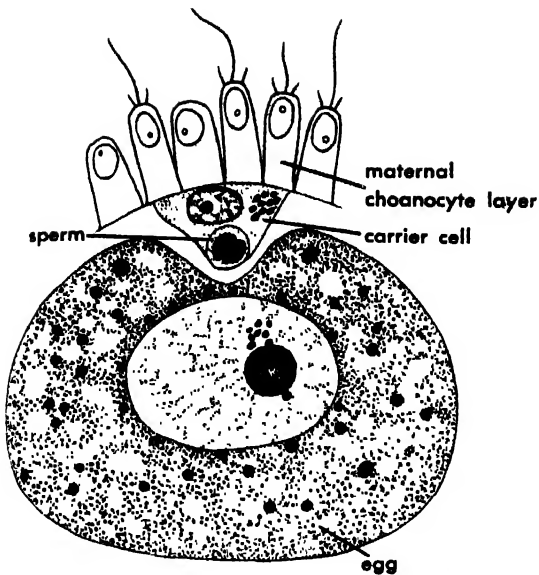


Fig. 7. Carrier cell transmitting sperm to egg. Calcareous sponge. (After Dubosq and Tuzet, 1937)

from the flagellated cells of the larva, from choanocytes, or from archaeocytes. The sperm enters the egg, not directly, but through the intermediary of a choanocyte which loses its collar and flagellum and migrates from the flagellated chamber to a position adjacent to the egg. This process has been observed in several species of *Calcarea* and *Demospongiae* and is probably of general occurrence in these groups. The sperm, which loses its tail and enlarges, is eventually transferred to the egg presumably following the fusion of the carrier cell with the egg. See FERTILIZATION.

Subsequent development of the embryo and larva varies somewhat in the several classes. The free-swimming flagellated larvae are either hollow amphiblastulae or solid stereogastrulae. See *CALCAREA*; *DEMOSPONGIAE*; *HEXACTINELLIDA*.

**Asexual reproduction.** Many fresh-water and marine sponges disintegrate with the onset of unfavorable environmental conditions, leaving behind reduction bodies which are compact masses of cells, chiefly amoebocytes, covered by an epidermal layer. In some shallow-water marine sponges the entire sponge remains alive during the winter months but changes internally into a mass of amoebocytes. Both reduction bodies and the partially dedifferentiated adult sponges can develop into functional colonies upon return of favorable conditions.

All fresh-water and many marine sponges produce asexual reproductive bodies, called gemmules, as part of their life cycle. Gemmules are formed of masses of archaeocytes laden with food reserves in the form of lipids and proteins. In fresh-water sponges the inner mass of cells is surrounded by a layer of columnar cells which secrete around the gemmule a double layer of spongin between which characteristic spicules are deposited. The gemmules of many species of marine sponges are also enclosed in a spongin coat, which may or may not be provided with spicules. The gemmules of

fresh-water sponges and a few marine species carry the species over periods of unfavorable environmental conditions such as drought or low temperatures, during which times the adult colonies degenerate. Germination of the gemmules occurs upon return of suitable conditions. In many marine sponges gemmules are formed at all seasons, and the adult colonies show no seasonal degeneration.

In a few species of *Demospongiae* and *Hexactinellida* asexually produced gemmules, lacking spongin coats, have been described as developing into flagellated larvae identical in structure to those formed as a result of sexual processes.

Budding is another common type of asexual reproduction in sponges. In *Tethya*, for example, groups of archaeocytes migrate to the tips of spic-

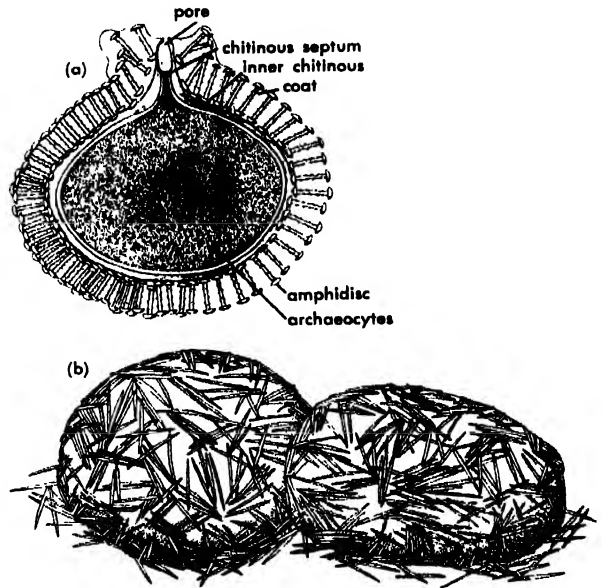


Fig. 8. Sponge gemmules. (a) Fresh-water sponge gemmule in optical section (after Evans, 1901). (b) Marine siliceous sponge gemmule, surface view (after Hartman, 1958).

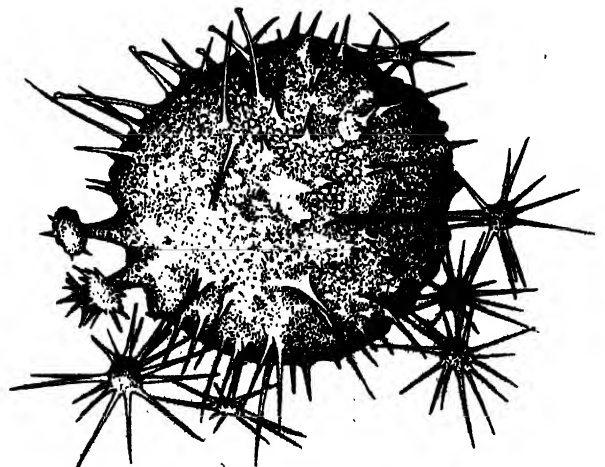


Fig. 9. Colony of *Tethya*, a marine siliceous sponge, with buds located at the tips of projecting spicule bundles. (From Edmondson, 1946)



Fig. 10. Sponge fishing off Florida. (After a mural by S. P. Glaser in the Peabody Museum of Natural History, Yale University)

ule bundles projecting from the surface. These buds fall off and form new individuals. See REPRODUCTION, ANIMAL.

**Regeneration.** Fragments cut from sponge colonies or pieces which break off accidentally in nature can reattach to the substrate and reconstitute functional individuals. Suspensions of sponge cells prepared by squeezing fragments of colonies through fine silk bolting cloth into sea water are also capable of reorganizing into functional sponges. All types of cells normally present in the adult sponge except scleroblasts have been identified in the cell suspensions. Reorganization of the cells to form functional colonies thus involves chiefly a migration of the cells to take up their appropriate positions in the reconstituted organism. The pinacocytes form a peripheral epithelium around the aggregated cells, and the various mesenchymal cells form the central mass. The collencytes are active in reforming the inhalant and exhalant canals, the latter of which become lined with pinacocytes. Choanocytes group together to reconstitute flagellated chambers between the inhalant and exhalant canals. Scleroblasts form anew from archaeocytes. Some workers have reported that collar cells dedifferentiate or are phagocytized by pinacocytes or amoebocytes and are formed anew during the reorganization of the sponge.

The initial aggregation of the dissociated cells to form masses results from random movement of archaeocytes, amoebocytes, and other cell types in-

cluding choanocytes which have been observed to put forth pseudopodia and which also move short distances by means of their flagella. As cells come into contact with one another they adhere, possibly as a result of antigen-antibodylike forces. The presence of homologous antibodies inhibits reaggregation. Neighboring cell aggregates often adhere as well if they come into contact with one another.

Mixed cell suspensions of two species of sponges of different colors may result in an initial intermixture of the cells of the two species, but these later sort out so that the cell aggregates are eventually composed exclusively or predominantly of cells of one species or the other.

**Commercial sponge fisheries.** Although plastic sponges now offer competition, there is still a demand for natural sponges for use by various artisans, for surgical purposes, and for cleaning automobiles. Commercially valuable sponges are harvested in the eastern part of the Mediterranean Sea, off the west coast of Florida and off the Florida Keys, in the West Indies, and to a limited extent off the Philippines. They are gathered by hooking or harpooning from a boat in shallow waters; by nude diving, a method used especially for exploiting cave populations in the Aegean Sea; by machine diving with the aid of diving suits attached by a life line to an air pump; and by dredging, a wasteful method prohibited on most sponge grounds.

Artificial propagation of sponges has proven feasible when serious efforts have been made to do so. In the Bahamas sponge cuttings planted on natural bottoms or on concrete disks in carefully selected areas yielded a harvest of 140,000 sponges during the years 1935-1939. Regrettably a disease, presumably caused by a fungus, spread through the West Indies in 1938-1939, put an end to the cultivation experiments, and killed a large percentage of the natural population.

Before World War II the Japanese had achieved considerable success in cultivating sponges in the Marshall and Caroline Islands. Their methods in-

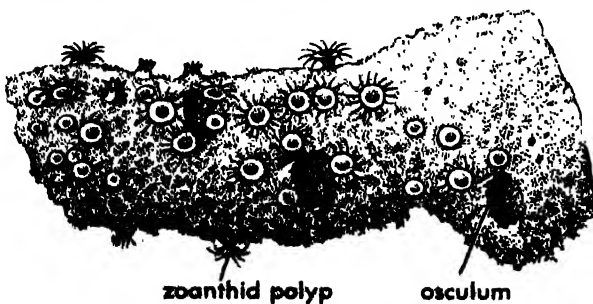


Fig. 11. Zoanthids living in a sponge.

cluded planting directly on concrete disks or stringing cuttings on aluminum wire anchored to rafts of various sorts or provided with floating buoys.

**Organisms associated with sponges.** Sponges are preyed upon regularly by chitons, snails, and nudibranchs, and occasionally by fish and probably other types of animals. The larvae of neuropteran insects of the family Sisyridae (spongilla flies) live in and feed upon colonies of fresh-water sponges. Sponges also play host to myriads of animals which live in the natural cavities of the sponge or form burrows of their own. Sea anemones, zoanthids, polychaetes, octopuses, copepods, barnacles, amphipods, shrimps, brittle stars, and fish live in sponge colonies, but the nature of their relationships with sponges has been little explored. The siliceous sponge *Suberites domunculus* always lives on shells occupied by hermit crabs and provides the crabs with a continually growing house. Certain crabs plant or hold pieces of sponge on their backs and thus blend into their environment. Other encrusting animals such as hydroids, bryozoans, and ascidians, which compete for space with sponges, often overgrow or are overgrown by sponge colonies in the struggle.

Zoochlorellae commonly live in the amoebocytes of fresh-water sponges and many shallow-water marine species. Because the algal cells fail to develop starch reserves when living in sponge cells, it seems likely that the sugars synthesized by the algae are utilized by the sponge cells. Dying algal cells are consumed by the amoebocytes. Apparently both the algae and the sponges can survive independently of each other. Young sponges are reinfected with algae from the water in which they live.

Certain marine sponges are regularly filled with filamentous green or coralline algae. The blue color of some sponges has been attributed to bacteria living in the cells. The significance of these relationships is unknown. See ANIMAL KINGDOM; ARCHAEOCYATHA; PARAZOA. [W.D.H.]

**Bibliography:** L. H. Hyman, *The Invertebrates*, vol. 1, 1940, vol. 5, 1959.

## Porifera fossils

The Porifera, or sponges, have a fossil record extending from the Cambrian to Recent times. More than a 1000 genera of fossil sponges have been described from the Paleozoic, Mesozoic, and Cenozoic eras. The preservation of fossil sponges is usually poor because the spicular skeleton usually becomes disjointed after death. Sponge spicules are commonly scattered through marine sedimentary strata, but their identification and assignment to established species, genera, families, and orders are usually impossible and certainly ill advised. Some forms with more solid skeletons, such as the lithistid Demospongia, dictyid Hyalospongia, and pharetrone Calcispongia left recognizable fossils.

**Geologic record.** According to M. W. de Laubenfels, the Cambrian and various succeeding periods yield fossil records which indicate abundance

of sponges comparable to that of the present. Lithistids were common in the Middle Silurian; Calcareas have a range from the Cambrian to the Recent, but were not common before the Devonian; and large numbers of Hyalospongia were also present in the Devonian. The Carboniferous cherts of Great Britain are largely composed of sponge spicules. Jurassic formations of Europe contain numerous lithistid Demospongia. The Cretaceous was a time of large numbers of Hyalospongia. The Cenozoic assemblages resemble those of today and mark a certain decline of numbers. In many cases the fossil sponges are associated with reef development. This is particularly true of the Silurian lithistids and the Permian lithistids and calcareous pharetrones. See ORGANIC REEF.

**Types.** Fossil sponges, like their living representatives, are divided into three classes: the Calcispongia (Calcareas), the Hyalospongia (Hexactinellida) and the Demospongia.

**The Calcispongia.** The Calcispongia (Calcareas) have calcareous spicules made of calcite or aragonite. The spicules are in the form of monaxons, triradiates, and quadriradiates. In the pharetrones the spicules are fused together into a more or less rigid network. The majority of fossil Calcispongia are pharetrones (order Pharetronida). In many cases the calcareous material of the fossil is replaced by secondary silica. The earliest known representative of the class appeared in the Cambrian but the group is represented, although poorly, until the Devonian.

**The Hyalospongia.** The Hyalospongia (Hexactinellida) have siliceous spicules made of opaline silica. The rays of the spicules are always at right angles to each other and number 4, 5, 6, and rarely 8 rays. The spicules are interlaced or fused at the tips producing an open and reasonably rigid meshwork. The stratigraphic range is from Cambrian to Recent. Excellently preserved glass sponges are known from the Devonian of New York State.

**The Demospongia.** The Demospongia have skeletons made of siliceous spicules, of spongin, or of siliceous spicules and spongin. It is rare that spicules of some kind are not present. The spicules can be monoaxial, tetraxial, or irregular. The vast majority of fossil Demospongia belong to the order Lithistida. These sponges have irregular and knobby siliceous spicules with commonly branching or expanded extremities. The adjacent spicules interlock producing a rigid stony skeleton. The thickened appearance of the spicules is the result of deposition of secondary silica on ordinary single or 4-rayed small spicules. The over-all shape of the sponge is often globular or ovoid and in some fossil forms outlines of body canals are preserved. The lithistids are reasonably well represented in Paleozoic rocks but become most abundant in the Jurassic and Cretaceous. See CALCAREA; DEMOSPONGIAE; HEXACTINELLIDA. [V.J.O.]

**Bibliography:** R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, Part E, Geol. Soc. Am.,

1955; V. J. Okulitch and S. J. Nelson, *Sponges of the Paleozoic*, Geol. Soc. Am. Mem. 67:763-770, 1957.

## Porocephalida

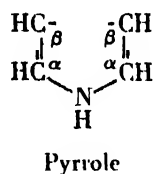
An order of the class Pentastomida in the phylum Arthropoda. The two families comprising this order are the Porocephalidae, with 5 subfamilies and 12 genera, and the Linguatulidae, with 2 genera. Distinctive features of this order include four-legged embryos, posterior location of the female genital pore, highly developed head and hook glands, and a long and winding ovary.

The Porocephalidae are cylindrical with club-shaped ends. They are provided with either simple hooks or double outer hooks. This family includes the greatest number of pentastomid species. The adult forms are parasitic in reptiles; the larval and nymphal forms infest mammals.

The Linguatulidae are flat in shape, with simple hooks in the adult state and binate hooks in the nymph. The first described species, *Linguatula serrata*, commonly occurs in the nasal cavities of dogs in Northern Europe. For this species, the larvae are approximately 5 millimeters (mm) long. The adult males measure 20 mm, and the females, 130 mm in length. See PENTASTOMIDA. [H.R.H.]

## Porphyrin

A class of red-pigmented compounds with a cyclic tetrapyrrolic structure in which the four pyrrole rings are joined through their  $\alpha$ -carbon atoms by



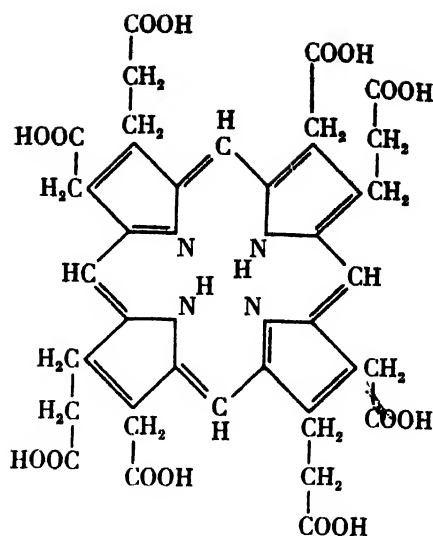
four methene bridges ( $=CH-$ ). The porphyrins form part of the active nucleus of chlorophylls *a* and *b*, hemoglobin, myohemoglobin, cytochromes, and the enzymes catalase and peroxidase. The parent substance is synthetic porphin in which the hydrogen (H) atoms in the eight  $\beta$ -positions on the pyrrole rings are unsubstituted. Naturally occurring porphyrins differ from porphin and from each other by various side chains in these eight  $\beta$ -positions. Some typical porphyrins are listed as follows:

1. Uroporphyrin occurs naturally and may be synthesized. The substituted groups are four carboxyethyl ( $-\text{CH}_2-\text{CH}_2-\text{COOH}$ ) and four carboxymethyl ( $-\text{CH}_2-\text{COOH}$ ).

2. Coproporphyrin occurs naturally and may be synthesized. The substituted groups are four carboxyethyl and four methyl ( $-\text{CH}_3$ ).

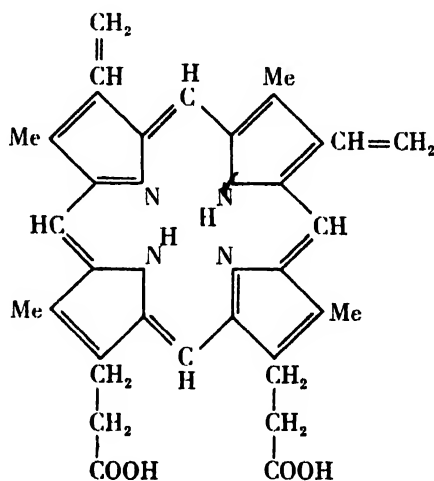
3. Etioporphyrin, a synthetic porphyrin, has four ethyl ( $-\text{CH}_2-\text{CH}_2-$ ) and four methyl groups.

4. Protoporphyrin occurs naturally and may be synthesized. The substituted groups are two car-



Uroporphyrin type I

boxyethyl, four methyl, and two vinyl ( $-\text{CH}=\text{CH}_2$ ).



Protoporphyrin type IX

5. Hematoporphyrin, a synthetic porphyrin, has two carboxyethyl, four methyl, and two hydroxyethyl ( $-\text{CH}_2-\text{CH}_2-\text{OH}$ ).

Porphyrins with two kinds of substituent groups, such as uro-, capro-, and etioporphyrin, have four structural isomers, known as types I-IV. Only types I and III are found in nature. The number of possible isomers for porphyrins with three different substituents, for example proto- and hematoporphyrin, is 15 (types I-XV). Protoporphyrin type IX has been identified in nature, in the free form and in heme which is the prosthetic group of hemoglobins and other heme-proteins. Protoporphyrin type IX corresponds to synthetic etioporphyrin type III.

Porphyrins dissolved in organic solvents and in dilute alkali have a typical absorption spectrum, exhibiting four bands in the visible range and a very strong Soret band in the near-ultraviolet. In

strong acid, the bands in the visible range are reduced to two.

Solutions of porphyrins in organic solvents or in mineral acids exhibit intense red fluorescence, with a single emission band in the 600 m $\mu$  range, while the exciting light may be of visible or near-ultraviolet (Soret band) wavelengths. For estimation of porphyrins, both spectrophotometric and fluorimetric methods are suitable. See CHLOROPHYLL; CYTOCHROME; ENZYME; HEMOGLOBIN. For biosynthesis see IRON METABOLISM. [R.S.C.]

## Porphyroblast

A relatively large crystal formed in a metamorphic rock. The presence of abundant porphyroblasts gives the rock a porphyroblastic texture. Minerals found commonly as porphyroblasts include biotite, garnet, chloritoid, staurolite, kyanite, sillimanite, andalusite, cordierite, and feldspar. Porphyroblasts are generally a few millimeters or centimeters across, but some attain a diameter of over 1 ft. They may be bounded by well-defined crystal faces, or their outlines may be highly irregular or ragged. Very commonly they are crowded with tiny grains of other minerals that occur in the rock.

Some porphyroblasts appear to have shoved aside the rock layers (foliation) in an attempt to provide room for growth. Others clearly transect the folia-

tion and appear to have replaced the rock. The presence of ghostlike traces of foliation, in the form of stringers and trails of mineral grains, passing uninterruptedly through a porphyroblast is further evidence of replacement.

Porphyroblasts have many features in common with phenocrysts but are to be distinguished from the latter by the fact that they have developed in solid rock in response to metamorphism. Most commonly they develop in schist and gneiss during the late stages of recrystallization. As the rock becomes reconstituted, certain components migrate to favored sites and combine there to develop the large crystals. See GNEISS; METAMORPHIC ROCKS; PHENOCRYST; SCHIST. [C.A.C.A.]

## Porphyry

An igneous rock characterized by porphyritic texture, in which large crystals (phenocrysts) are enclosed in a matrix of very fine-grained to aphanitic (not visibly crystalline) material.

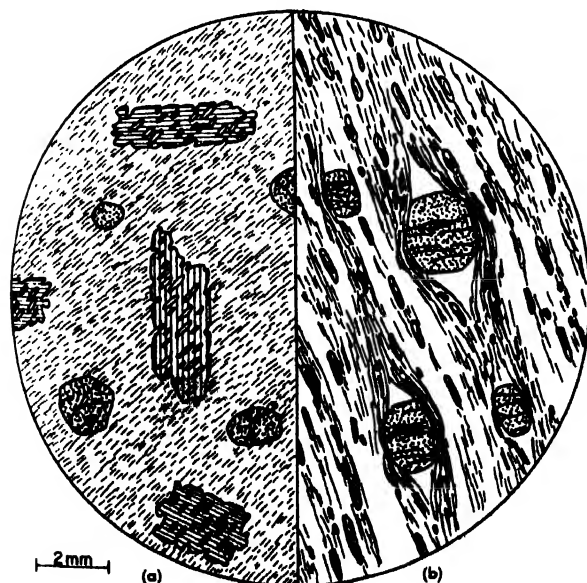
Porphyries are generally distinguished from other porphyritic rocks by their abundance of phenocrysts and by their occurrence in small intrusive bodies (dikes and sills) formed at shallow depth within the earth. In this sense porphyries are hypabyssal rocks.

Compositionally, porphyries range widely, but varieties may be distinguished by prefixing to the term the common rock name which the porphyry most closely resembles (such as granite porphyry, rhyolite porphyry, syenite porphyry, and trachyte porphyry). See IGNEOUS ROCKS.

Porphyries are gradational to plutonic rocks on the one hand and to volcanic rocks on the other. In the granite clan, for example, six porphyritic types may be recognized: (1) porphyritic granite, (2) granite porphyry, (3) rhyolite porphyry, (4) porphyritic rhyolite, (5) vitrophyre, and (6) porphyritic obsidian or porphyritic pitchstone.

A rock of granitic composition with abundant large phenocrysts of quartz and alkali feldspar in a very fine-grained matrix of similar composition is a porphyry, or more specifically a granite porphyry. Granite porphyry passes into porphyritic granite as the grain size of the matrix increases and abundance of phenocrysts decreases. Thus the rock becomes a granite, or more specifically a porphyritic granite. Granite porphyry passes into rhyolite porphyry as the grain size of the matrix and abundance of phenocrysts decrease. The principal distinction between rhyolite porphyry and porphyritic rhyolite is the mode of occurrence. Rhyolite porphyry is intrusive; porphyritic rhyolite is extrusive. Porphyritic rocks with a glass matrix are known as vitrophyres. With decrease in number of phenocrysts, vitrophyre passes into porphyritic obsidian or porphyritic pitchstone.

The phenocrysts of porphyries consist largely of quartz and feldspar. Quartz occurs as well-formed (euhedral) hexagonal bipyramids, which in thin section under the microscope exhibit a diamond-



(a) Porphyroblastic mica schist. Large crystals (porphyroblasts) of biotite mica formed late and replaced the rock except for the enclosed grains (quartz) aligned parallel to the foliation. Smaller porphyroblasts of garnet replaced the rock completely. (b) Porphyroblastic quartz-mica schist. Large porphyroblasts of garnet have grown by spreading apart the mica-rich layers of the schist. The elongate grains (gray) are flakes of biotite mica. Finer flakes of muscovite mica are widespread and give a pronounced schistosity. They are closely packed where crowded by the garnet crystals. Elongate grains of iron and titanium oxide (black) are also aligned.



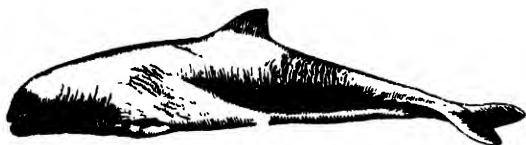
square, or hexagonal outline. Individual phenocrysts may show more or less rounding or resorption with deep embayments. Alkali feldspar is usually euhedral sanidine, orthoclase, or microperthite. Plagioclase occurs more in association with phenocrysts of hornblende or other dark-colored (mafic) minerals. Porphyritic rocks with predominantly mafic phenocrysts (olivine, pyroxene, amphibole, and biotite) are commonly classed as lamprophyres. See LAMPROPHYRE.

Outside the United States, it is common to further restrict the term porphyry to those rocks in which the feldspar phenocrysts are principally alkali feldspar. Rocks with dominantly plagioclase phenocrysts are called porphyrites.

Porphyries occur as marginal phases of medium-sized, igneous bodies (stocks, laccoliths) or as apophyses (offshoots) projecting from such bodies into the surrounding rocks. They are also abundant as dikes cutting compositionally equivalent plutonic rock or as dikes, sills, and laccoliths injected into the adjacent older rocks. [C.A.C.A.]

## Porpoise

- Any of several small whales of the family Phocaenidae, found in all the oceans of the world except the polar waters. The porpoises, like other whales, have no hindlimbs. The forelimbs are reduced to small paddles; the caudal region is modified into a pair of lateral flukes; the neck region is shortened; and the eyes are small. Porpoises have many small teeth, set all along the jaws; the dorsal



The common porpoise, *Phocaena phocaena*; length to 6 ft. (From E. L. Palmer, *Fieldbook of Natural History*, McGraw-Hill, 1949)

fin is low or lacking; the head is blunt and rounded, without a beak. Porpoises attain a maximal length of 8 ft, and weigh not more than 120 lb. The harbor porpoise, *Phocaena phocaena*, is common along both coasts of the United States. It is 4–6 ft long, black on the back, with pink sides and a white belly. The Pacific race is considered by some students to be a separate species, *P. vomerina*. See MAMMALIA; WHALE. [J.D.B.]

## Portland cement

A hydraulic cement consisting mainly of calcined silicates of calcium. Portland cement mixed with water, sand, and gravel (or aggregate) or other substances is a common material of construction. See CONCRETE.

**Manufacture.** The illustration shows the manufacture of portland cement. The raw materials are principally calcareous materials such as limestone,

marl, chalk, or shells and argillaceous materials such as clay, shale or iron blast-furnace slag. The chemical limitations in the specifications are such that the proportions of calcium oxide, silica, alumina, and ferric oxide must be maintained within narrowly defined limits; and other constituents, such as magnesia and alkalies, must not exceed specified limits. These restrictions necessitate at times the introduction of other materials, such as a high-calcium limestone, sandstone, or iron ore. To produce white portland cement, materials of very low ferric oxide content must be used. The raw materials are blended and ground to a fineness approximately the same as cement. The raw materials may be ground dry (the dry process) or in water (the wet process). The pulverized mixture is then burned in large rotary kilns at 2500–2800°F to produce portland cement clinker. The fuel may be coal, oil, or gas.

Calcium sulfate (gypsum) is added in a quantity ranging from 4 to 8% by weight of the clinker. It serves to regulate the setting time, strength, and other properties of cement-water mixtures. The clinker and gypsum are then ground to such a fineness that about 90% will pass through a no. 200 sieve, that is, a wire or cloth mesh having 200 openings per square inch.

**Types of cement.** Heat is liberated during the process of hydration and hardening of cement-water mixtures. It is desirable to minimize this heat liberation in mass concrete, such as large dams, so as to reduce the thermal stresses and cracking that may occur as the exterior surface cools while the interior of the mass is at a higher temperature. Some types of cement are susceptible to chemical reactions with sulfate waters which can cause expansion and deterioration of mortars and concretes. The heat of hydration and sulfate resistance are controlled by adjusting the chemical composition of the cement. High early strength is attained in part by composition but largely by finer grinding of the cement.

The American Society for Testing Materials specifies the chemical and physical requirements for five types of portland cement and describes the purpose of each.

Type I is for use in general concrete construction when the special properties specified for types II, III, IV, and V are not required.

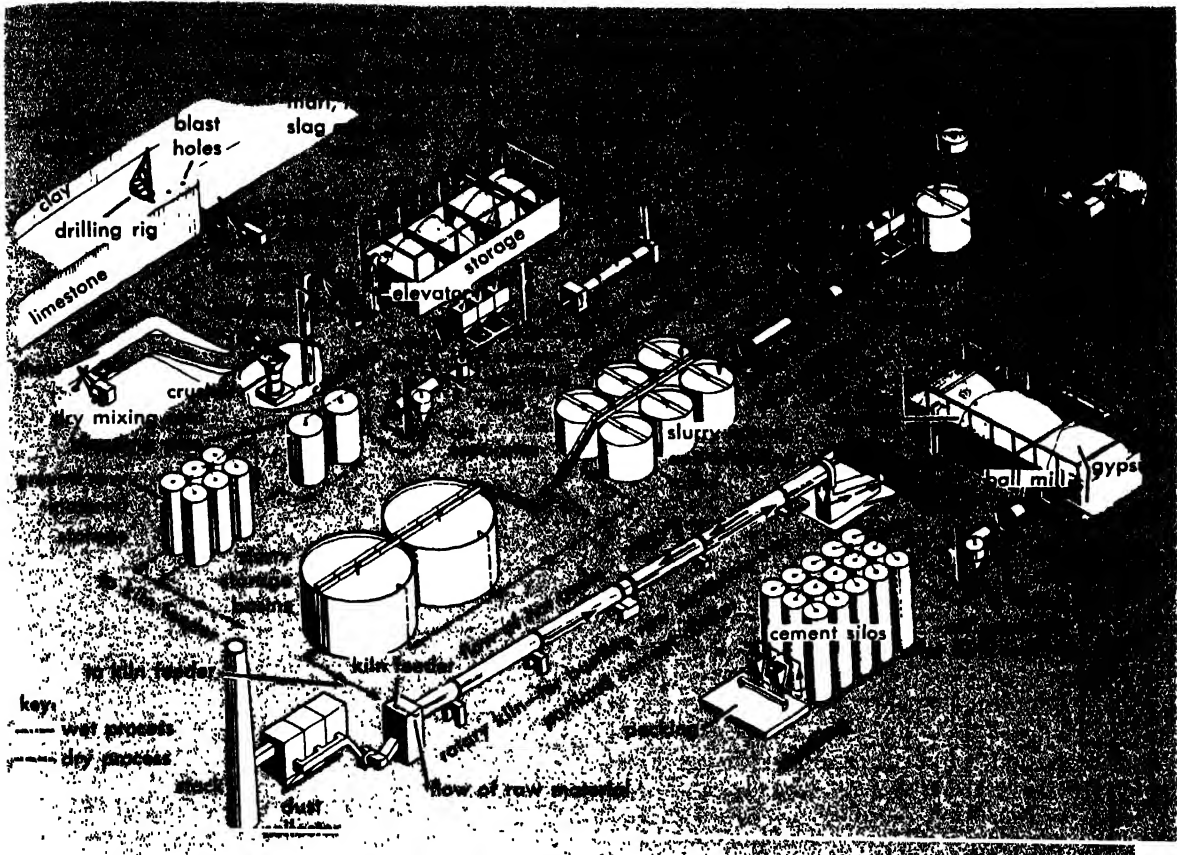
Type II is for use in general concrete construction exposed to moderate sulfate action, or where moderate heat of hydration is required.

Type III is for use when high early strength is required.

Type IV is for use when a low heat of hydration is required.

Type V is for use when high sulfate resistance is required.

There are also a number of modifications of portland cement or mixtures containing portland cement. See AIR-ENTRAINING PORTLAND CEMENT; GEMENT.



Flow chart (simplified) showing the manufacture of portland cement from quarrying to final shipment.

Portland cement is used in many different types of construction. The use pattern is approximately as follows: highways, 20%; nonresidential construction, 20%; residential building, 15%; military construction, 10%; public utilities, 10%; sewer and water works, 8%; other uses, 17%.

[W.L.E.]

### Portuguese man-of-war

A large, floating, highly complex, colonial, marine animal of the class Hydrozoa, phylum Coelenterata. The most prominent feature of the Portuguese man-of-war is the large float, or central body, often 8–10 in. long, bearing a sail-like crest. This float is beautifully iridescent, showing blues, pinks, purple, and carmine. Beneath this float are structures of three distinct types, some nutritive, others feelers, and still others reproductive. The tentacles carry numerous stinging cells, called nematocysts, capable of inflicting a painful, but rarely fatal, injury. The tentacles of a large specimen may be over 60 ft long. The animal feeds primarily on fish which are first subdued by the nematocysts. Although most abundant in the Gulf of Mexico coastal waters, it occurs in numbers in the Gulf Stream of the Atlantic Ocean, and frequently drifts northward for some distance.

A small species of fish of the genus *Nomeus* lives among the tentacles of the Portuguese man-



The Portuguese man-of-war, *Physalia pelagica*; length of body to 10 in. (From J. G. Wood, *Popular Natural History*, Porter and Coates)

of-war with apparent immunity. The fish seems to be accepted and protected by the man-of-war as a lure to attract other larger fishes. The *Nomeus*, in turn, is thought to feed on scraps left over from the kills made by the man-of-war. This is frequently cited as a prime example of symbiosis. Some workers think that *Nomeus* is immune to the stings of the Portuguese man-of-war because it feeds on the animals caught in its tentacles. See HYDROZOA.

[J.D.B.]

## Positron

An elementary particle with mass equal to that of the electron, and positive charge equal in magnitude to the electron's negative charge. The positron is thus the antiparticle (charge-conjugate particle) to the electron (see ELECTRON; ELEMENTARY PARTICLE). Its existence was predicted by P. A. M. Dirac (see QUANTUM THEORY, RELATIVISTIC). It was first observed by C. D. Anderson in 1932. The positron has the same spin and statistics as the electron. Positrons, like electrons, appear as decay products of many heavier particles; electron-positron pairs are produced by high-energy photons in matter. See PAIR PRODUCTION (ELECTRON-POSITRON).

A positron is, in itself, stable, but cannot exist indefinitely in the presence of matter, for it will ultimately collide with an electron. The two particles will be annihilated as a result of this collision, and photons will be created. However, a positron can first become bound to an electron to form a short-lived "atom" termed positronium. See POSITRONIUM.

The virtual production of electron-positron pairs by an electromagnetic field produces a polarization of the vacuum. This results in effects such as the scattering of light by light and modification of the electrostatic Coulomb field at short distances. See QUANTUM ELECTRODYNAMICS. [C.J.G.]

**Bibliography:** W. Heitler, *The Quantum Theory of Radiation*, 3d ed., 1954; J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons*, 1955.

## Positronium

The bound state of an electron and a positron. Positronium was discovered by studies of the so-called annihilation radiation from positrons stopped in gases. It is formed in a collision between a positron and a gas atom which results in the capture of an atomic electron by the positron. The positron is the antiparticle to the electron and hence has an inertial mass equal to that of the electron, a positive charge equal in magnitude to the electron's charge, and a spin of  $\hbar/2$ , where  $\hbar$  is Planck's constant  $h$  divided by  $2\pi$ . See POSITRON.

Positronium is of particular interest because it is the two-body system to which quantum electrodynamics is applicable, and its study has served as an important confirmation of the theory of quantum electrodynamics. See QUANTUM ELECTRODYNAMICS.

No states of positronium other than the ground  $n = 1$  state ( $n = 1, 2, 3, \dots$ , being the principal quantum number) have yet been found. Studies of positron annihilation in solids and liquids indicate that a perturbed form of positronium exists under certain conditions.

**Energy levels.** The approximate energy levels of positronium can be calculated from the Schrödinger equation with the nonrelativistic Hamiltonian,  $H_0$ :

$$H_0 = \frac{p_1^2}{2m} + \frac{p_2^2}{2m} - \frac{e^2}{r} \quad (1)$$

in which  $p_1(p_2)$  is the electron (positron) linear momentum,  $m$  is the mass of the electron or positron,  $-e$  is the charge of the electron, and  $r$  is the distance between the positron and the electron (see QUANTUM THEORY, NONRELATIVISTIC). The energy levels of the bound states are given by

$$W_n = -\frac{\pi^2 m e^4}{h^2 n^2} = -\frac{r_{\nu p}}{n^2} \quad (2)$$

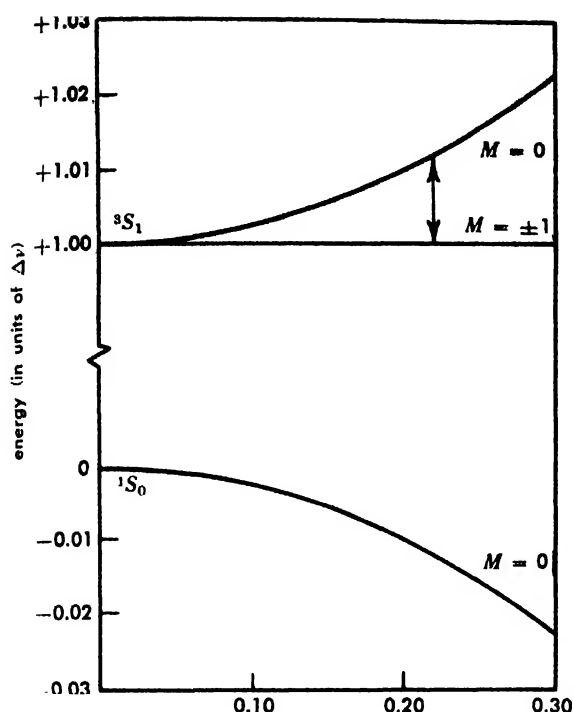
The quantity  $r_{\nu p}$  is defined by Eq. (2) as the Rydberg constant for positronium (see RYDBERG CONSTANT). The binding energies  $W_n$  of positronium are one-half the corresponding binding energies of the hydrogen atom (if the proton-to-electron mass ratio is considered infinite). In particular, the ionization energy of positronium (the binding energy of the ground  $n = 1$  state) is 6.8 ev.

Fine structure to the energy levels of positronium as given by Eq. (2) of the order  $\alpha^2 r_{\nu p}$  [ $\alpha = e^2/\hbar c \cong 1/137$  is called the fine structure constant] arises from relativistic effects including the electron and positron spin magnetic moments and from the interaction with the electromagnetic field which causes electron-positron pair annihilation. Since the electron and positron intrinsic spin angular momenta are  $\frac{1}{2}$  in units of  $\hbar$ , the total spin angular momentum quantum number  $S$  of positronium can be either 0 (singlet state, parapositronium) or 1 (triplet state, orthopositronium). For each  $n$  value, positronium can exist in either a singlet or a triplet state. The orbital angular momentum quantum number  $L$  can assume the values  $L = 0, 1, \dots, n-1$ . In particular, the ground  $n = 1$  state of positronium is split into two levels,  $^1S_0$  and  $^3S_1$ , which are separated in energy by the amount

$$W(^3S_1) - W(^1S_0) = \alpha^2 r_{\nu p} \left[ \frac{7}{3} - \frac{2\alpha}{\pi} \left( \frac{16}{9} + \ln 2 \right) \right] \quad (3)$$

The term of order  $\alpha^3 r_{\nu p}$  arises from virtual quantum electrodynamic processes. This energy separation, often called the hyperfine structure of the ground state of positronium, corresponds to a frequency difference  $\Delta\nu$  of  $2.0337 \times 10^6$  Mc/sec. See FINE STRUCTURE (SPECTRAL LINES); HYPERFINE STRUCTURE.

The dependence of the energy levels of positronium on an external magnetic field (Zeeman ef-



Zeeman energy levels of positronium in its ground  $n = 1$  state. The quantity  $\Delta\nu$  is the hyperfine structure separation between the  $^3S_1$  and the  $^1S_0$  states of positronium at zero static magnetic field. The  $M$  values designate the magnetic substates.  $x = 2g_{s1}\mu_0 H / (h\Delta\nu)$ .

fect) can be determined from the Hamiltonian term

$$H_H = \mu_0 g_{s1} s_1 \cdot H + \mu_0 g_{l1} l_1 \cdot H + \mu_0 g_{s2} s_2 \cdot H + \mu_0 g_{l2} l_2 \cdot H \quad (4)$$

in which 1(2) refers to the electron (positron),  $\mu_0$  is the Bohr magneton ( $= e\hbar/2mc$ ),  $g_l$  is the orbital  $g$ -value ( $= 1$ ),  $g_{s1}$  is the electron spin  $g$ -value [ $= 2(1 + \alpha/2\pi - 0.328 \alpha^2/\pi^2)$ ],  $g_{s2} = -g_{s1}$ ,  $l_{1(2)}$  is the electron (positron) orbital angular momentum,  $s_{1(2)}$  is the electron (positron) spin angular momentum, and  $H$  is the external magnetic field intensity. Positronium has no permanent magnetic moment, but there can be a magnetic moment induced by the external magnetic field, and hence an energy level can depend on  $H^2$  or higher powers of  $H$ . The energy level diagram for the ground state of positronium in a magnetic field is shown in the figure. From measurements of the frequency of the Zeeman transition  $\Delta M = \pm 1$  between the magnetic sublevels of the  $^3S_1$  state, the hyperfine structure interval  $\Delta\nu$  has been determined to be

$$\Delta\nu = (2.0333 \pm 0.0004) \times 10^5 \text{ Mc/sec}$$

This experimental value agrees with the theoretical value, and the agreement constitutes the principal test of the quantum electrodynamics of the two-body problem.

**Decay.** Positronium is an unstable atom and annihilates with the emission of photons. From its ground  $^1S_0$  state a positronium atom at rest decays into two  $\gamma$ -rays each having an energy of  $mc^2$  ( $\sim 510$  kev) with a decay rate of  $8.03 \times 10^9 \text{ sec}^{-1}$ ; from its ground  $^3S_1$  state a positronium atom at rest decays into three  $\gamma$ -rays whose energies total  $2mc^2$  with a decay rate of  $7.21 \times 10^9 \text{ sec}^{-1}$ . [v.w.h.]

**Bibliography:** S. DeBenedetti and H. C. Corben, Positronium, *Ann. Rev. Nuclear Sci.*, 4:191-218, 1954; M. Deutsch, Annihilation of positrons, *Progr. in Nuclear Phys.*, 3:131-138, 1953; S. Fluegge (ed.), *Handbuch der Physik*, vol. 35, 1956.

## Postglacial vegetation and climate

The Pleistocene or Glacial Epoch, at present estimated as the last 1,000,000 years, is less remarkable for radical evolutionary changes in plant life than for repeated shifts of position because of climatic changes and the resulting ice movements. The influence of these changes is known to have extended far beyond the limits of glaciation, particularly in mountainous regions, whose belts or zones of vegetation migrated up and down in response to climate. The extent of horizontal shifting far beyond the ice borders is not yet resolved. See GLACIAL EPOCH; PLEISTOCENE.

**Plant migration and glacial stages.** The present distribution of vegetation, certainly outside the tropics, expresses the readjustment that followed the last major ice advance, known as the Wisconsin age in North America. The retreat of the Wisconsin ice from the extreme position it occupied in southern Ohio some 18,000 years ago was pulsating in character, being marked by alternating periods or substages of retreat and readvance. Twelve thousand years ago it had retreated, then readvanced to Port Huron, Michigan. Some 2000 years later it had again retreated and readvanced to the north shore of Lake Erie. From this Valdres stage it then began a rapid retreat to its present position, with a possible pause in Canada some 6000 years ago.

The European equivalent of the Valdres is the Fennoscandian, marked by the northernmost conspicuous moraines, dated between 10,000 and 11,000 years ago. Subsequent time is arbitrarily considered postglacial in Europe. In America it appears more convenient to speak of postglaciation time with reference to specific localities, since, for example, Columbus, Ohio, and Long Island, New York, were ice-free long before Detroit, Michigan, and Hartford, Connecticut, and have had a correspondingly longer vegetation history.

The earlier evidence of plant migration and hence of inferred climatic change was derived from macroscopic plant remains and from disjunct or detached species and communities of living plants now found outside their characteristic range. Classic examples are the communities of bog plants, northern in affinities, that occur within the decidu-

ous forest region. These are properly considered as remnants left behind from a colder time.

The record was complicated, however, by the presence of disjuncts appropriate to warmer and drier conditions than now exist, such as prairie inclusions in the deciduous forest regions of North America, and Mediterranean plants in northwestern Europe. Furthermore, the European peat beds contained strata indicating that the time of glacial retreat had been marked by fluctuations in moisture and temperature and was not a simple, gradual warming down to the present.

To reconstruct an orderly history requires the study of representative plant remains in close stratigraphic sequence. These requirements are met by the fact that pollen grains are well preserved in lake sediments, both inorganic and organic. The statistical study of such evidence, begun by L. von Post in Sweden in 1916, is the most substantial source of present information. See PALYNOLOGY.

Broadly speaking, the results obtained confirm the hypothesis advanced by the Norwegian Axel Blytt in 1876 on the basis of his study of Scandinavian peat beds. There are, however, many local modifications due to soils, topography, rates of plant migration, and masking of minor changes by climatic extremes. Blytt's ideas were, for these and other reasons, the source of prolonged controversy. He divided postglacial time into the following phases: Pre-Boreal, cold and wet; Boreal, cool and dry; Atlantic, warm and moist; Sub-Boreal, warm and dry; Sub-Atlantic, cooler and moist. Recent work suggests that the Boreal in northern Europe was warmer than the Sub-Boreal. The reverse seems to be true in North America but further study is needed.

Until recently the only method of dating these changes was by means of banded sediments or varves, and by estimates of the rate of sedimentation, in particular of peat formation. Both methods considerably lessened the previously accepted time span, and gave results of an order of magnitude now confirmed by the precise physical measurements of radiocarbon change.

**Vegetation and climatic phases.** Four general principles emerge from the data now at hand: (1) The general pattern of climate following ice retreat is that of an irregularly sinusoidal graph

so far as moisture is concerned. (2) This pattern occurs against a background of temperature gradients. (3) The maximum number of vegetation and climatic phases occur near the glacial boundary, the number decreasing toward the present areas of remnant ice. (4) Moist phases appear to coincide with time of ice accumulation and advance, dry phases with ice wastage and retreat. Whether or not these phases are reversed within the tropics is now under discussion.

Also under discussion is the extent to which major zones of vegetation—such as tundra, conifer forest, deciduous forest—were displaced beyond the ice border. Work now being done in the arid and mountainous Southwest, based on sediments in old lake basins, shows clearly the downward migration of high altitude woodland and forest during glacial times and its subsequent replacement by semidesert or desert vegetation.

The accompanying table although much simplified and omitting certain interesting fluctuations of the past 2000 years, illustrates the general character of postglaciation changes. See PALEOBOTANY; PLANT GEOGRAPHY; VEGETATION ZONES. [P.B.S.]

*Bibliography:* E. L. Braun, *Deciduous Forests of Eastern North America*, 1950; Pierre Dansereau, *Biogeography*, 1957; H. Godwin, *The History of the British Flora*, 1956; P. B. Sears, Xerothermic theory, *Botan. Rev.*, 8(10):708-736, 1942.

## Postulate

In a formal deductive system a proposition accepted without proof, from which other propositions are deduced by the conventional methods of formal logic. There is a certain arbitrariness as to which propositions are to be treated as postulates, because when certain proved propositions are treated as postulates, other propositions which were originally postulates often become proved propositions.

The question of objective truth does not arise in connection with a postulate, although the term postulate is sometimes loosely used in connection with tentative assumptions with regard to matters of fact. In strict usage, postulate is nearly equivalent to axiom, although axiom is often loosely used to denote a truth supposed to be self-evident. See LOGIC. [P.W.B.]

Postglaciation phases in North America

Phase	Moisture	Temperature	Ice	New England	Southern Ohio	Northern Ohio	Alaska
8	—	W	R				Tundra-spruce
7	+	W±	A	Hemlock-chestnut	Beech-maple	Beech-maple	(Glaciation)
6	—	W+	R	Oak-hickory	Oak-hickory	Oak-hickory	Spruce
5	+	W±	A	Oak-hemlock	Beech-hemlock	Beech-hemlock	—
4	—	?	R	Pine	Pine	Pine	—
3	+	C	A	Spruce	Spruce	Spruce	—
2	—	C	R	Tundra	Pine	(Glaciation)	—
1	+	C	A	(Glaciation)	Spruce	—	—

Symbols. Phase numbers unofficial, for convenience only. Moisture signs: (+) moist phase, (—) dry phase. Temperature: C—cold, W—warm. Ice: A—advance, R—retreat.

## Posture, regulation of

The contraction of muscles serves not only to move the joints in locomotion but also to maintain the joints in fixed positions during the assumption of stationary attitudes or postures (see MOTOR SYSTEMS). A simple example of the role of the neuromuscular system in maintaining posture is seen when a man maintains a standing position. When the body is in the upright position, gravitational forces tend to cause the joints to flex and the limbs to collapse under the weight of the body. This is prevented by the contraction of the extensor muscles of the lower limb so that gravitational forces are counteracted and the upright position is maintained. The muscular contraction required for erect standing is maintained by reflex action and requires no conscious effort or attention; interruption of the nerves supplying the leg muscles destroys not only the ability to walk but also the ability to stand (see REFLEX, UNCONDITIONED). The reflex mechanisms regulating posture are highly flexible so that passive or active changes in position of the body cause autonomic changes in the pattern of muscle contraction, which result in stances or attitudes appropriate to the new or changing orientation of body. The reflexes mediating these reactions are known as postural reflexes.

**Decerebrate rigidity.** The spinal reflex mechanism responsible for standing may be studied in intact animals but is best seen in a decerebrate preparation, an animal in which the brain stem has been completely transected between the superior and inferior colliculi to produce total isolation of the lower brain stem, medulla, and spinal cord from the rostral (anterior) brain structures. The cat survives such radical surgery surprisingly well; despite the occurrence of certain neurological deficits, the vital functions of respiration and circulation are adequately maintained by the remaining isolated neural structures. A striking feature of the decerebrate preparation is an exaggerated posture first described by C. S. Sherrington in 1898. The limbs assume a position of rigid extension, and actively resist passive flexion of the joints. Once forcibly flexed, the joints are easily extended. Palpation of the muscles reveals that the rigidity is maintained by active contraction of the extensor muscles of the limbs.

Extensor rigidity is sufficient to support the weight of the body, and when the animal is placed on its feet, it stands rigidly if it is lightly supported from the side to prevent toppling. Muscles other than those of the limbs are similarly affected; the head is held erect with the chin up, the jaw is clamped tightly shut and actively resists efforts to open it, and the tail often stands erect. The muscles displaying rigidity are the antigravity muscles, that is, the muscles which in the normal upright stance oppose the gravitational forces, tending to cause collapse of the body. Decerebration in the sloth, an animal which resists gravity with its

flexor muscles as it hangs upside down from branches, causes a flexor rigidity.

Decerebrate rigidity is thus a caricature of standing. The exaggeration is due to the release of spinal motoneurons from inhibitory impulses which originate in suprasegmental structures and normally descend through those pathways which have been interrupted by the brain stem transection. The role of these suprasegmental structures is discussed below. Decerebrate rigidity is a reflex that requires segmental afferent input to maintain the discharge of the motoneurons supplying the affected antigravity muscles. When the dorsal spinal roots conveying sensory impulses from a rigid limb to the spinal cord are divided, rigidity disappears and the limb becomes flaccid. Similarly, ventral root section abolishes rigidity by interrupting the efferent part of the reflex arc.

The origin of the afferent impulses which maintain decerebrate rigidity is not in receptors of skin, of joints, or of tendons because it persists unabated after removal of skin, and after local anesthetics, which render receptors insensitive, are injected into joints and tendons. The responsible receptors are in the muscles themselves and the adequate stimulus for their excitation is stretch of the muscle; because they thus register events occurring in the body itself they are classed as proprioceptors as opposed to exteroceptors (for example, retinal and tactile receptors) which react to external events. The impulses generated in the stretch receptors of antigravity muscles by gravitational forces (which tend to cause the limb to flex and thus stretch the extensor muscles) are transmitted to the spinal cord over afferent fibers of the dorsal roots. The "stretch-sensitive" afferents make direct connections with the motoneurons supplying the stretched muscle and initiate in them an efferent discharge which elicits muscular contraction opposing elongation of the muscle and thus maintains upright posture. Further, the efferent discharge is graded so that reflex muscular contraction is, through a considerable range of stretching forces, sufficient to maintain the muscle at fixed length and thus to prevent collapse of the limb.

For example, in unburdened standing, only the weight of the body tends to buckle the knees and stretch the knee extensors. This mild degree of stretch excites only the stretch receptors of lowest threshold, and these receptors discharge at relatively low rates. Consequently the evoked efferent discharge and the resultant reflex contraction is precisely sufficient to keep the joint straight. If a heavy load is placed upon the shoulders, the additional stretching force causes active stretch receptors to discharge impulses at higher rates and also elicits discharges in those higher threshold receptors which were previously silent (recruitment). The increased afferent discharge activates more motoneurons and drives them at higher rates so that the resultant reflex contraction is increased sufficiently to offset the newly imposed burden. The



stretch, or myotatic, reflex, as it is called, provides an automatically regulated mechanism for upright posture, which functions effectively regardless of varying demands placed on the limb musculature.

The stretch reflex is responsible for muscle tone, that is, the resistance of the muscle to passive elongation. When the stretch reflex is absent, the muscle is hypotonic or flaccid; when stretch reflexes are exaggerated as in decerebrate preparations, the muscles are hypertonic or spastic.

A quantitative estimate of the sensitivity and effectiveness of the stretch reflex can be obtained from experiments employing the fall-table of Sherrington, the important feature of which is a movable top which can be lowered for measured distances. The leg of an experimental animal is fixed rigidly to a stand on the table, and a muscle is dissected free and attached to a tension-recording device mounted on a stand independent of the movable table top. When the table top is lowered, the muscle is stretched and the tension developed in the muscle recorded. Part of the tension is caused by the elasticity of the muscle; this moiety can be determined by denervating the muscle and repeating the stretch. The difference in tension developed in the innervated and denervated muscle gives a measure of the tension caused by active reflex contraction. In the quadriceps (knee extensor) muscle of the cat a stretch of 8.0 mm gives rise to an active reflex tension of 3–3.5 kg.

**Stretch receptor and its control.** The sense organ of muscle which mediates the myotatic reflex is the muscle spindle. This organ consists of six or seven small specialized muscle fibers, called intrafusal fibers. They are enclosed in a connective tissue capsule, the ends of which attach to the tendon or to the connective tissue surrounding the ordinary muscle fibers (extrafusal fibers). Afferent nerve fibers penetrate the capsule and terminate in a helical arrangement (annulospiral ending) or a bushy ending (flower-spray ending) on the equatorial region of the intrafusal fibers. The equatorial region of the intrafusal fibers is noncontractile but the two ends or poles of the fibers are contractile and receive motor innervation from small-diameter efferent fibers which penetrate the capsule along with the sensory fibers. The spindles, which lie in parallel with the extrafusal fibers, are subjected to stretch when the muscle is elongated, and the resultant distortion of the nerve endings excites them to discharge impulses. However, if the extrafusal fibers contract and the muscle is allowed to shorten, the intrafusal fibers become slack and the endings cease firing. The slack can be taken up if the small motor fibers supplying the contractile poles of the intrafusal muscle fibers are activated. These small motor fibers (3–6  $\mu$  in diameter) are histologically and functionally distinct from the large motor fibers (9–13  $\mu$  in diameter) which innervate the extrafusal fibers. The former, which comprise about 30% of the fibers in the ventral root of the spinal cord, are devoted exclusively to the innervation of intrafusal fibers, and are therefore

called fusimotor fibers. Fusimotor excitation develops no measurable tension in the muscle, because the contractile tension of the small intrafusal fibers is negligible compared to that of the extrafusal fibers. Instead, the function of the fusimotor fibers appears to be that of a sensitivity control for the spindles. When fusimotor activity is great, the spindles are kept taut and slight muscle stretches elicit a high rate of afferent discharge. In the absence of fusimotor activity the spindles are slack, and greater muscle stretches are required to establish a comparable discharge. The stretch-reflex mechanism therefore includes a loop, from muscle spindle to motoneuron causing extrafusal contraction and spindle silence and one from sense organs to fusimotoneurons. The afferent paths which drive the fusimotoneurons are not entirely understood. Part of the drive is through segmental pathways; noxious stimulation of the skin is particularly effective in increasing fusimotor discharge and secondarily increasing the spindle afferent discharge to stretch. Certain descending paths from suprasegmental structures also exert a marked influence on fusimotoneurons.

**Regulation of the stretch reflex.** In the decerebrate preparation the stretch reflex is abnormally sensitive, with the result that slight stretches induce inordinately large efferent discharges. This hypersensitivity is ascribed to the interruption by the intercollicular section of pathways originating rostral (anterior) to the section, and exerting, directly or indirectly, an inhibitory influence on the segmental stretch-reflex mechanism. The origin of those pathways is not entirely clear. In man and higher primates spasticity, a syndrome similar to decerebrate rigidity, follows lesions of the rostral precentral cortex. Probably the inhibitory systems originate from several supracollicular structures.

Removal of an inhibitory system alone will not cause exaggerated reflex response, any more than removing the brakes from a stalled car will make it go. The question of what drives the motoneuron released from supracollicular inhibition is considered next. The segmental afferents from the stretch receptors provide part of the drive but suprasegmental structures also contribute, because spinal transection abolishes decerebrate rigidity as readily as dorsal root section. Between the spinal cord and the colliculi there must be a group of neurons or a center which facilitates the segmental stretch reflex arc. Experiments indicate that facilitatory impulses originate in the medulla and reach the spinal level partly by way of the vestibulospinal tract and partly by the reticulospinal paths. At the spinal level, these impulses sensitize the stretch reflex by converging with segmental afferents from stretch receptors on motoneurons and by driving the fusimotoneurons, thus increasing the sensitivity of the spindles to stretch and augmenting the segmental afferent input to the motoneurons.

**Lengthening reaction.** When the limb of a decerebrate cat is forcibly stretched, resistance develops and increases throughout the initial part of the

bending but at some point suddenly disappears so that the muscle elongates readily. This lengthening reaction is reflex in origin. The receptor is the Golgi tendon organ and its discharge causes inhibition of the motoneurons of the stretched muscle. Complete inhibition occurs only with strong stretches because the Golgi tendon organs have higher thresholds than muscle spindle endings. The lengthening reaction protects the muscle against overload, preventing the development of injurious tensions with excessive stretches.

**Tonic neck reflexes.** The segmental stretch reflex provides for upright posture but a variety of sensory inputs alters its stereotyped pattern and permits a flexible variety of postures or attitudes. When the head of the decerebrate cat is bent forward on the chest, the forelimbs flex and rigidity increases in the hindlimbs. The resulting attitude is that of a cat looking under a bed. Conversely, dorsiflexion of the head increases forelimb rigidity and causes flexion of the hindlimbs; this is the postural pattern of a cat looking up at the top of a bed and ready to spring. The receptors mediating these attitudinal reflexes are proprioceptors in the joints of the neck.

The rotation of the head around the long axis of the neck elicits another postural change; extensor tone increases in the limbs, both fore and hind, toward which the chin is turned. On the opposite side extensor tone diminishes. The responsible receptors are again proprioceptors in the neck joints; section of the cervical dorsal roots abolishes the reflex pattern.

**Tonic labyrinthine reflexes.** Even after the tonic neck reflexes are eliminated by cervical dorsal root section, changing the animal's position in space alters the extensor tone of the limbs. Extensor tone is maximal in all four limbs when the animal is horizontally supine with the snout at an angle of 45° above the horizontal plane. Rotation of the animal around the horizontal axis progressively reduces extensor tone, a minimum being reached in the prone position with the snout tilted 45° below the horizontal plane. The responsible receptor is the otolith organ of the labyrinth of the ear; the reflex alterations are static and do not depend upon acceleration of the head. See EAR.

**Righting reflexes.** When dropped through space or displaced from the upright position, the intact cat rapidly reorients its body so that it lands on its four feet or attains an upright position. This righting ability is lost in the decerebrate cat which stands only when held upright. Righting is not entirely dependent upon visual cues because it occurs in blinded animals or after complete removal of the cerebral cortex. Reflex righting is accomplished by the following mechanisms.

**Labyrinthine reflexes.** Disorientation of the head is detected by the otolith receptors of the labyrinth; the reflex discharge is to the neck muscles so that the head assumes a horizontal position. The reflex is a static reflex; that is, it does not depend upon acceleration of the head.

**Body-righting reflexes.** When a blinded labyrinthectomized animal is placed on its side, the head turns and assumes the horizontal position. The reflex is initiated by the asymmetric stimulation of the body surface by the weight of the body.

**Neck-righting reflexes.** When the head has been righted by the labyrinthine and body-righting reflexes, the resultant twisting of the neck excites proprioceptors which reflexively affect trunk and limb musculature so that the body assumes a horizontal position.

**Body reflexes acting on body.** Asymmetric stimulation of the body surface in an animal lying on one side causes the body to right itself even when the head is held in the lateral position.

**Visual righting reflexes.** In the absence of all the foregoing reflex mechanisms, visual stimuli cause righting of the head and body. Optical or visual righting is dependent on the cerebral cortex.

**Placing and hopping reactions.** These are postural reflexes which, like the visual righting reflexes, depend upon the integrity of the cerebral cortex. Placing reactions ensure that the feet are placed under the body in the proper position for standing; they may be visual or nonvisual. When an animal is lowered by hand toward a visible support, the limbs seek the support. If a blindfolded animal is similarly lowered, contact of the feet with the support sufficient to excite tactile endings or muscle proprioceptors initiates nonvisual placing of the feet. Hopping reactions are elicited when the body is displaced horizontally with the feet trailing on the floor; the feet execute a series of hops so that the legs remain in a supporting position vertical to the body. Hopping reactions are both visual and nonvisual; in the latter instance, muscle proprioceptors initiate the reflex.

**Labyrinthine acceleratory reflexes.** The role of labyrinthine receptors in static postural reflexes has already been described. The adequate stimulus for other labyrinthine reflexes is neither position nor velocity of displacement but rather acceleration of the head. It was once thought that the maculae of the utricle and saccule of the inner ear were responsible for the static reflexes, whereas the cristae of the semicircular canals were the receptive mechanism underlying the acceleratory reflexes. Recent evidence suggests that the dichotomy is not this sharp and that both utricle and semicircular canals contain receptors mediating both static and acceleratory responses.

The simplest acceleratory reflex is elicited by linear acceleration, as for example, sudden lowering, head down, of a blindfolded animal causes the forelimbs to extend and the toes to spread. This response, which is lacking in the labyrinthectomized animal, supports the animal in landing from a jump. Angular acceleration elicits labyrinthine reflex alterations of the musculature of neck, limbs, trunk, and eyes. The response of the eyes is called nystagmus. As the head rotates, the eyes swing slowly in the opposite direction (slow phase) to maintain fixation of the gaze until, reaching a

maximum deviation, they swing rapidly back in the direction of rotation to the forward-looking position (quick phase); the process repeated over and over causes an oscillation of the eyes. When the head reaches constant velocity, nystagmus disappears, only to reappear when rotation ceases; in the latter instance, however, the quick phase is in the direction opposite to the original direction of rotation.

Nystagmus may be horizontal, vertical, or rotatory depending upon the position of the head during angular acceleration. When the head is bent forward 30° the horizontal semicircular canals are in the plane of rotation and the nystagmus is horizontal. When the head is flexed 90° onto the shoulder, the vertical canals are in the plane of rotation and nystagmus is vertical. With the head bent forward 120° or backward 60° rotation causes rotatory nystagmus. See MUSCLE; MUSCLE (BIOPHYSICS); NERVOUS SYSTEM; PSYCHOLOGY, PHYSIOLOGICAL AND EXPERIMENTAL. [H.D.P.; T.C.R.]

## Potassium

With an atomic number of 19, and an atomic weight of 39.1, potassium, K, stands in the middle of the alkali metal family, below sodium and above rubidium, in group Ia of the periodic table of the elements. It is a lightweight, soft, low-melting, reactive metal. In 1807, Sir Humphry Davy isolated metallic potassium, by electrolysis, for the first time. It is very similar to sodium in its behavior in metallic form, and its uses are limited as a consequence of the availability of low-cost sodium in large volume.

iron, calcium, and sodium are more abundant. Sea water contains 380 parts per million, making potassium the sixth most plentiful element in solution, being exceeded only by chlorine, sodium, magnesium, sulfur, and calcium.

Potassium compounds are found in the historically important deposits at Stassfurt in Germany, which consist of sylvite (KCl) and carnallite ( $\text{MgCl}_2 \cdot \text{KCl}$ ). In the United States, extensive potassium deposits containing sylvite and polyhalite ( $2\text{CaSO}_4 \cdot \text{K}_2\text{SO}_4 \cdot 2\text{H}_2\text{O}$ ) are located at Searles Lake, California, and Carlsbad, New Mexico. In addition, potash salts are found in France, Spain, Poland, the Soviet Union, and in the Dead Sea.

**Metallurgical extraction.** Commercial potassium chloride is melted in a gas-fired melt pot and is fed to the exchange column (see illustration) in the commercial manufacture of potassium metal by thermochemical means. The molten potassium chloride flows down over steel Raschig rings in the packed column. It is met by ascending sodium vapors coming from a gas-fired reboiler. An equilibrium is set up between the two, giving sodium chloride and potassium metal as the products. The sodium chloride formed is continuously withdrawn at the base of the apparatus. The column operating conditions may be varied to give practically pure potassium metal as an overhead product or to vaporize sodium along with the potassium to give sodium-potassium (NaK) alloys of varying compositions as products. Potassium metal of over 99.5% purity can be produced continuously.

Unlike lithium and sodium, which are produced by electrolysis, potassium reacts with carbon electrodes, and also can form an explosive carbonyl in electrolysis (or in thermochemical methods using carbon reduction). Therefore, the thermochemical route using the reaction between metallic sodium and potassium chloride has proven most practical and economic.

**Physical properties.** The physical properties of potassium metal are summarized in the table.

Ia

IIa

IIIa IVa Va VIa VIIa 0

IIb IVb Vb VIb VIIb VIII Ib IIb

19 K

lanthanum series

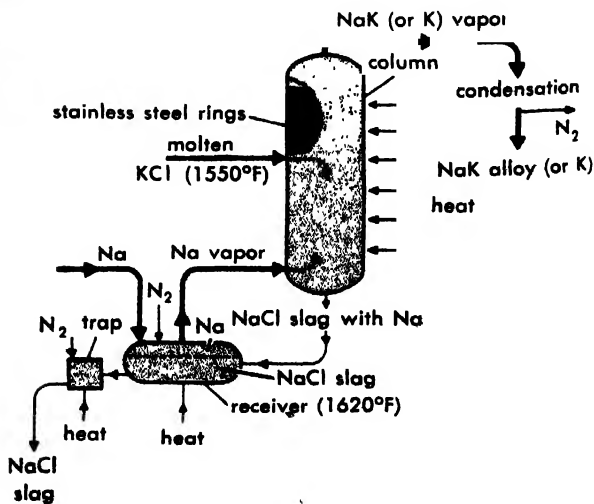
actinium series

**Uses.** Potassium chloride finds its main use in fertilizer mixtures. It also serves as the raw material for the manufacture of other potassium compounds.

Potassium hydroxide is used in the manufacture of liquid soaps, and potassium carbonate in making soft soaps. Potassium carbonate is also an important raw material for the glass industry.

Potassium nitrate is used in matches, in pyrotechnics, and in similar items which require an oxidizing agent.

**Occurrence.** Potassium is a very abundant element, ranking seventh among all the elements in the earth's crust, 2.59% of which is potassium in combined form. Only oxygen, silicon, aluminum,



### Potassium extraction.

## Physical properties of potassium metal

Property	Temperature		Metric (scientific) units	British (engineering) units
	°C	°F		
Density	100	212	0.819 g/cm <sup>3</sup>	51.1 lb/ft <sup>3</sup>
	400	752	0.747 g/cm <sup>3</sup>	46.7 lb/ft <sup>3</sup>
	700	1292	0.676 g/cm <sup>3</sup>	42.2 lb/ft <sup>3</sup>
Melting point	63.7	147		
Boiling point	760	1400		
Heat of fusion	63.7	147	14.6 cal/g	26.3 Btu/lb
Heat of vaporization	63.7	1400	496 cal/g	893 Btu/lb
Viscosity	70	158	5.15 millipoises	6.5 kinetic units
	400	752	2.58 millipoises	3.5 kinetic units
	800	1472	1.36 millipoises	2 kinetic units
Vapor pressure	342	648	1 mm	0.019 lb/in. <sup>2</sup>
	696	1285	400 mm	7.75 lb/in. <sup>2</sup>
Thermal conductivity	200	392	0.017 cal/(sec)(cm <sup>2</sup> )(cm)(°C)	26.0 Btu/(hr)(ft <sup>2</sup> )(°F)
	400	752	0.09 cal/(sec)(cm <sup>2</sup> )(cm)(°C)	21.7 Btu/(hr)(ft <sup>2</sup> )(°F)
Heat capacity	200	392	0.19 cal/(g)(°C)	0.19 Btu/(lb)(°F)
	800	1472	0.19 cal/(g)(°C)	0.19 Btu/(lb)(°F)
Electrical resistivity	150	302	18.7 microhm-cm	
	300	572	28.2 microhm-cm	
Surface tension	100–150		About 80 dynes/cm	

**Chemical properties.** Potassium is even more reactive than sodium. It reacts vigorously with the oxygen in air to form the monoxide, K<sub>2</sub>O, and the peroxide, K<sub>2</sub>O<sub>2</sub>. In the presence of excess oxygen, it readily forms the superoxide, KO<sub>2</sub> (formerly believed to be K<sub>2</sub>O<sub>4</sub>).

Potassium does not react with nitrogen to form a nitride, even at elevated temperatures. With hydrogen, potassium reacts slowly at 200°C and rapidly at 350–400°C. It forms the least stable hydride of all the alkali metals.

The reaction between potassium and water or ice is violent, even at temperatures as low as –100°C. The hydrogen evolved is usually ignited in reaction at room temperature. Reactions with aqueous acids are even more violent and verge on being explosive.

Instead of forming the carbide with carbon, potassium forms a rather indefinite solid solution with the potassium atoms interposed between the layers of the graphite lattice.

Potassium reacts vigorously with the halogens. Lithium and sodium react only superficially with liquid bromine, but potassium detonates in contact with it. Potassium ignites in the reaction with iodine, also.

The reaction of potassium with ammonia gives potassium amide, KNH<sub>2</sub>, and hydrogen. Potassium differs from sodium in that an explosive carbonyl is formed when potassium reacts directly with carbon monoxide.

Potassium reacts with many organic compounds, but not with saturated aliphatic hydrocarbons. With some aromatic hydrocarbons, metalation occurs, giving organopotassium compounds. With acetylene, potassium acetylides are formed.

Potassium reacts with alcohols to form alkoxides and hydrogen. Most reactions of potassium with organic carbonyl compounds are very similar to those of sodium. In the form of NaK alloy, potassium is a very effective catalyst for the transesterification

reaction involved in the commercial modification of lard.

**Availability.** Potassium metal is available in one grade of 99+ % purity, with sodium as the major impurity. The annual production is about 50,000 lb/year. The price of the metal varies widely with the quantity ordered: 1 5-lb lots cost \$4.75 per lb in 1958 and 100,000 300,000-lb lots cost \$1.00 per lb. Potassium salts are generally more expensive than the corresponding sodium salts. Potassium hydroxide and potassium carbonate sell for about 9¢ per lb (as of 1958) compared to 2–3¢ per lb for the corresponding sodium compounds.

**Handling.** Handling of potassium metal is much the same as that of sodium metal, with two major exceptions. First, the formation of the superoxide, KO<sub>2</sub>, causes difficulties because it can react vigorously with hydrocarbons and other organic matter. Second, potassium is generally more reactive than sodium. Potassium forms an explosive carbonyl with carbon monoxide, and the metal detonates in contact with bromine.

Generally sodium, potassium, and the sodium-potassium (NaK) alloys are considered to be in the same general class of reactivity, allowing for the chemical differences outlined above and for the liquid (and hence more reactive) nature of the NaK alloys over a wide composition range. See SODIUM.

**Principal compounds.** Potassium chloride, KCl, is the most important potassium compound. It is not only the form in which potassium is often found in nature, but it is the form in which potash is used as a fertilizer.

Potassium hydroxide, KOH, is also known as caustic potash. It is usually made by the electrolysis of aqueous solutions of potassium chloride.

Potassium carbonate, K<sub>2</sub>CO<sub>3</sub>, is made from potassium hydroxide and carbon dioxide. It cannot be made by the Solvay process used for sodium car-

bonate because potassium bicarbonate is too soluble in ammonium chloride solution.

Potassium nitrate,  $\text{KNO}_3$ , is made by fractional crystallization of an aqueous solution containing sodium nitrate and potassium chloride.

**Analytical methods.** As in the case of sodium, the high water solubility of most potassium compounds complicates the analytical determination of potassium. Qualitative detection is usually made by means of the violet potassium flame; the sodium flame, which is usually present as well, is masked by viewing the flame through a cobalt-glass filter.

Gravimetric determination of potassium can be made using sodium triphenylboron, sodium perchlorate, or other reagents. Rubidium and cesium interfere in most of these gravimetric methods when they are present. See ALKALI METALS. [M.SI.]

**Bibliography:** Am. Chem. Soc., *Handling and Uses of the Alkali Metals*, Advances in Chem. Ser., vol. 19, 1957; C. B. Jackson, *Liquid Metals Handbook, Sodium-NaK Supplement*, 3d ed., 1955; R. N. Lyon (ed.), *Liquid Metals Handbook*, 2d ed., Navexos P-733 (rev.), 1954.

## Potato, Irish

The white potato, *Solanum tuberosum*, of the plant order Tubiflorales, grown in cool climates. The Irish potato is the world's leading vegetable crop, with an annual production of more than 8,000,000,000 bu and ranks with wheat and rice among the world's most important foods.

Spaniards invading South America found the white potato under cultivation high in the Andes Mountains of Peru and Bolivia, the region believed to be its center of origin, and took it back to Spain in 1538. It was introduced into Ireland about 1586. History indicates that the potato was brought to Virginia from the West Indies about 1621. It was transported from Ireland by Scotch-Irish immigrants to Londonderry (now called Derry), New Hampshire, in 1719, and was given the name Irish potato. See TUBIFLORALES.

**Characteristics.** This herbaceous annual has round or angular, aerial, green or pigmented stems which have axillary branches. The stems vary from  $\frac{1}{2}$  to  $1\frac{1}{4}$  in. in diameter and may grow erect to a height of 3–4 ft, or they may be procumbent (creeping).

The leaves, which are pinnately compound, consist of a large terminal leaflet subtended by three or four pairs of large petiolated leaflets borne laterally on a rachis; the leaves are arranged spirally on the stem.

Fibrous roots arise on the underground stem in groups of three just above the nodes. The roots grow about 24 in. laterally and about 18 in. deep under cultivation in humid regions, but the depth may reach 3 ft or more under drier conditions.

The flowers occur in cymose inflorescences. Each flower has a white, pinkish, or purplish five-lobed corolla, and the anthers of the five stamens borne

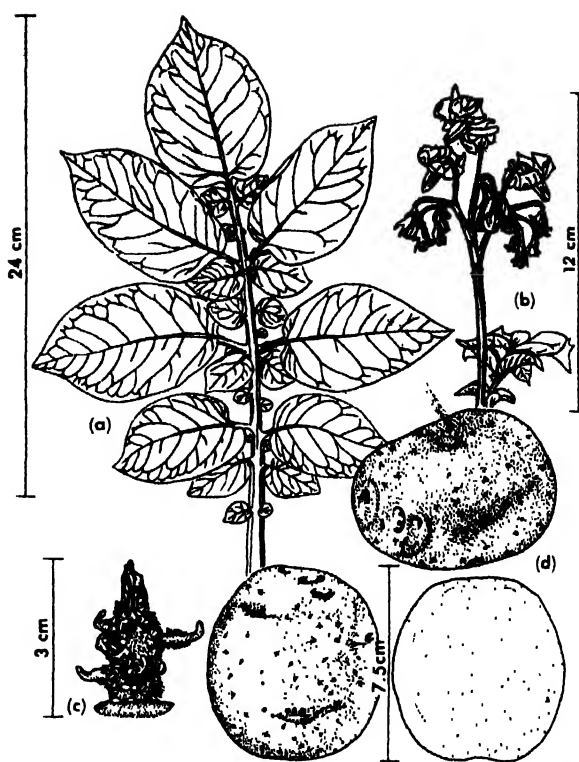


Fig. 1. Irish potato (Katahdin variety). (a) Pinnately compound leaf. (b) Flowers. (c) New plant developing from a tuber section ("seed" piece). (d) Tubers showing "eyes" containing buds. (Dutch Potato Atlas, H. Veenman and Zonen, Wageningen, Netherlands)

on the corolla tube are orange, lemon, or sometimes greenish yellow (Fig. 1).

Under cool conditions, varieties with fertile pollen may develop fruit balls which are green berry-like structures  $\frac{3}{4}$ – $1\frac{1}{4}$  in. in diameter and contain numerous kidney-shaped seeds (Fig. 2). These are sometimes mistaken for young tomatoes. Seeds are used experimentally in the production of new varieties. Desirable varieties, however, are propagated vegetatively by means of tuber sections which possess buds. These are often referred to as "seed" pieces.

Tubers are formed on the tips of or are sessile on underground lateral stems or rhizomes commonly called stolons. The tuber is a shortened, thickened, underground stem having nodes and internodes, the position of the nodes being indicated by "eyes" which are leaf scars containing lateral branches with axillary buds and very short internodes. Several sprouts may develop from one eye. Depending on the variety, tubers vary in shape from round to oblong or oval, in smoothness of skin, in depth of eyes, and in skin color from light yellowish-brown to red. Varieties also differ in time of maturity, in resistance to diseases and insects, in adaptation to different growing conditions, and in yield and marketable quality of tubers, including cooking and processing characteristics. Depending on growing conditions, marketable-sized



Fig. 2. Irish potato plant showing fruits. (USDA)

tubers over the 2-in. minimum diameter are produced within 2 months from date of planting in early varieties or within 4-5 months in later varieties.

**Varieties.** Since the organization in 1929 of the National Potato Breeding Program, many new varieties have been distributed, and some of these are displacing the older ones. New varieties are resistant to two or more diseases, are high-yielding, and are generally of good market and cooking quality. The 1957 United States certified seed-potato production report lists 13 old varieties introduced between 1850 and 1900 and 41 new ones released since 1932. The Katahdin variety, released in 1932, is now the most widely grown late variety. Following Katahdin in order of seed-potato production are Russet Burbank, Red Pontiac, Cobbler, Kennebec, White Rose, Chippewa, Red LaSota, Cherokee, Sebago, Red McClure, Triumph, Green Mountain, and Russet Rural. These varieties represented 93% of that year's certified seed potatoes (tubers cut into sections and used for reproduction).

The potato is produced commercially in nearly every state, with concentrated production in parts of Maine, Idaho, California, New York, Minnesota, North Dakota, and Colorado (Fig. 3). Since 1948

the annual production in the United States has ranged from 327,000,000 to 432,000,000 bu (196,000,000-259,000,000 cwt). During the same period the total annual farm value has varied from about \$300,000,000 to \$685,000,000.

**Chemical composition.** The chemical composition of the Irish potato varies greatly by variety, growing conditions including climate, soil, and fertilization, and the temperature and length of time in storage. Potatoes contain 75-85% water, 12-18% starch, about 2% protein, 1% inorganic compounds, about 0.5% sugar (in freshly harvested tubers), 0.5% organic acids (mainly citric), and 0.4% crude fiber. The Irish potato is a good source of vitamin C; 1 lb of boiled newly harvested potatoes contains the minimum daily adult requirement for this vitamin. However, the vitamin C content decreases with storage, especially at temperatures below 50°F.

**Use.** Annual consumption of potatoes is now about 100 lb per capita. A downward trend in consumption of unprocessed potatoes seems to have been stopped during the late 1950s, mainly because of increased use of processed products such as potato chips, frozen potato products, soups, and dehydrated mashed potatoes. Consumption of processed potatoes increased from 1.9 lb per capita in 1940 to 23.4 lb in 1956. In 1956, 45,000,000 bu was processed as potato chips, 10 times the amount prepared as chips in 1940, and 9,000,000 bu was prepared as frozen French fries compared with none in 1940. Processors demand potatoes high in solid-matter for potato chips and for frozen prepared products. In 1957 approximately 5,000,000 bu was commercially prepeeled as a service to restaurants and other large users.

Starch is made of cull and lower-grade potatoes by manufacturing plants in Maine and Idaho and on Long Island, in New York. The amount processed into starch has varied from less than 5,000,000 bu in 1951 to over 28,000,000 bu in 1956. Culls and lower-grade potatoes are also used for livestock

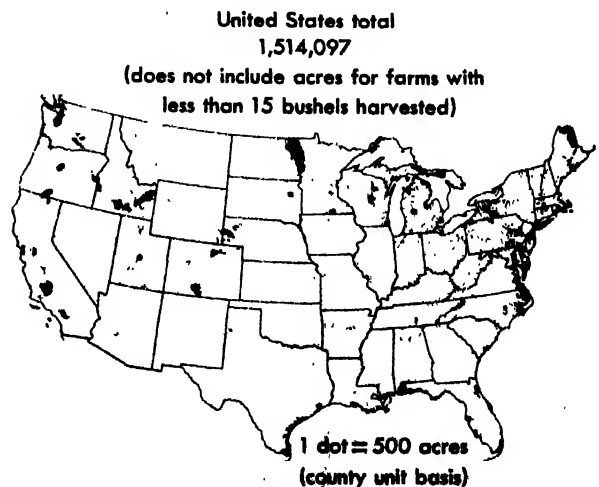


Fig. 3. Irish potato acreage in the United States for one year. (Bureau of the Census)



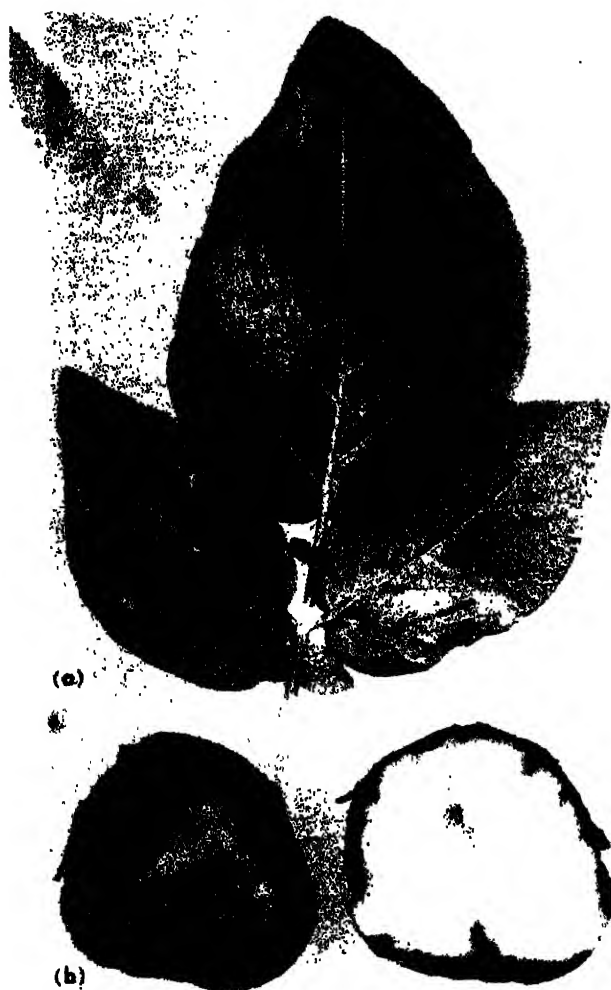


Fig. 4. Late blight of Irish potato, *Phytophthora infestans*. (a) Foliar infection. (b) Tuber infection. (USDA)

feed. See VEGETABLE GROWING; see also POTATO, SWEET. [A.H.]

**Irish potato diseases.** The potato is subject to numerous destructive diseases caused by fungi, bacteria, viruses, and nematodes. Potatoes may also be seriously injured by unfavorable growing or storage conditions.

**Fungus diseases.** The most destructive single disease is late blight caused by the fungus *Phytophthora infestans*, which was introduced into Europe after the potato had become established as a major food crop (Fig. 4). A succession of epidemics, starting about 1845, caused tremendous damage, especially in Ireland, where loss of the potato crop resulted in the starvation of 1,000,000 people and the emigration of 1,500,000 others. Late blight still causes important losses of potatoes in most of the important potato-growing areas, and makes control programs necessary.

**Less important fungus diseases of potato foliage** are early blight and certain rusts and powdery mildews. Some fungi, especially *Rhizoctonia*, *Verticillium*, *Sclerotinia*, and *Fusarium*, attack the stem

primarily, causing rots, cankers, or wilts. Tubers in the field are subject to attack by many fungi which cause such diseases as black wart, powdery scab, common scab, silver scurf, leak, rhizoctonia disease, pink rot, and fusarium tuber rot.

**Bacterial diseases.** Similar bacteria cause black-leg of the stem (*Erwinia atroseptica*) and bacterial soft rot of the tuber (*Erwinia carotovora*). The latter is extremely destructive in storage. Other destructive bacterial diseases are the ring and brown rots, both of which cause wilt or death of growing plants and rot or deterioration of tubers.

**Virus diseases.** As a group, viruses are probably the most important and insidious pathogens of the potato. Mild mosaic, rugose mosaic, leaf roll, latent mosaic, spindle tuber, calico, yellow dwarf, witches'-broom, haywire, and purple-top wilt are descriptive names of important virus diseases. Although most viruses do not kill the plant outright, they reduce the yield and quality of the tubers. Because most viruses are carried inside the tubers, the planting of certified disease-free tuber sections is the principal means of control.

**Nematode diseases.** Nematodes (small eelworms) cause root knot and golden nematode, tuber-rot nematode, and meadow nematode diseases. Nematodes often persist in the soil for many years, and, once established, they are difficult to control.

**Mechanical injury.** There also are many kinds of nonparasitic disease, such as blackheart, hollow heart, stem-end necrosis, and sprain. These injuries are caused by excessive heat or cold in the field or in storage, by nutritional disorders, by an excess or lack of water, and by poor ventilation in storage. See NEMATODA; PLANT DISEASE; PLANT VIRUS. [H.D.T.]

## Potato, sweet

The fleshy root of the plant *Ipomoea batatas*. The sweet potato was mentioned as being grown in Virginia as early as 1648. In 1930 the selection of outstanding strains of the Porto Rico variety, which was introduced into Florida in 1908, was begun in Louisiana, and the best strain, Unit I Porto Rico, was released in 1934. In 1937 new techniques were developed for inducing the sweet potato to bloom and set seed. This stimulated a surge of research on breeding for higher yield, greater nutritional value, better shape, storage ability, market and canning quality, greater disease resistance, and new food products and industrial uses for the sweet potato throughout the southern states and from New Jersey to California. In 1957 the crop in the United States was valued at \$71,427,000. In Louisiana, the leading commercial state for production of both the canned and fresh products, the annual value of the crop varies from \$15,000,000 to \$20,000,000, depending on seasonal conditions and demand.

**Types.** There are two principal types of sweet potato, the kind erroneously called yam and the Jersey type (Fig. 1). The chief difference between the two is that in cooking or baking the yam, much of



Fig. 1. Sweet potatoes. (a) Porto Rico, a yam variety. (b) Big-stem Jersey type. (USDA)

the starch is broken down into simple sugars (glucose and fructose) and an intermediate product, dextrin. This gives it a moist, syrupy consistency somewhat sweeter than that of the dry (Jersey) type. On cooking, the sugar in the dry type remains as sucrose.

The yam is produced largely in the southern states; however, because of the breeding of more widely adapted varieties, it is now being grown farther north. The Jersey sweet potato is grown largely along the eastern shore of Virginia, Maryland, Delaware, and New Jersey, and also in Iowa and Kansas (Fig. 2).

The total consumption of sweet potatoes in the United States, like that of white potatoes, rice, and

other high-carbohydrate foods, is now somewhat lower than in former years. Nevertheless, the value of the sweet-potato industry in Louisiana is increasing at the rate of about \$500,000 of market value each year. With the development of new varieties, the northern limits of sweet-potato production have been extended to Canada and northern United States. This has come about by breeding early and better varieties. In Michigan a number of baby-food manufacturers are growing this crop for processing.

**Breeding.** The principal objectives in sweet-potato breeding are higher nutritional values, including higher vitamin and mineral contents; increased yield; greater disease resistance; and wider adaptation. Louisiana, Oklahoma, Georgia, North and South Carolina, and more recently California have instigated breeding programs. The U.S. Department of Agriculture and several states are joining in the testing and evaluation of the newer seedlings and varieties. For example, each year in Louisiana about 20,000 seedlings are grown and studied. Men from all parts of the world are being trained in breeding, production, and handling the crop. The better varieties are sent to practically all the countries of the tropical, subtropical, and most of the temperate zones.

**Processing.** In the United States the sweet potato is canned extensively in Louisiana and along the Eastern Shore, and its popularity is increasing each year. The frozen product has also appeared on the market. The sweet-potato chip is another new

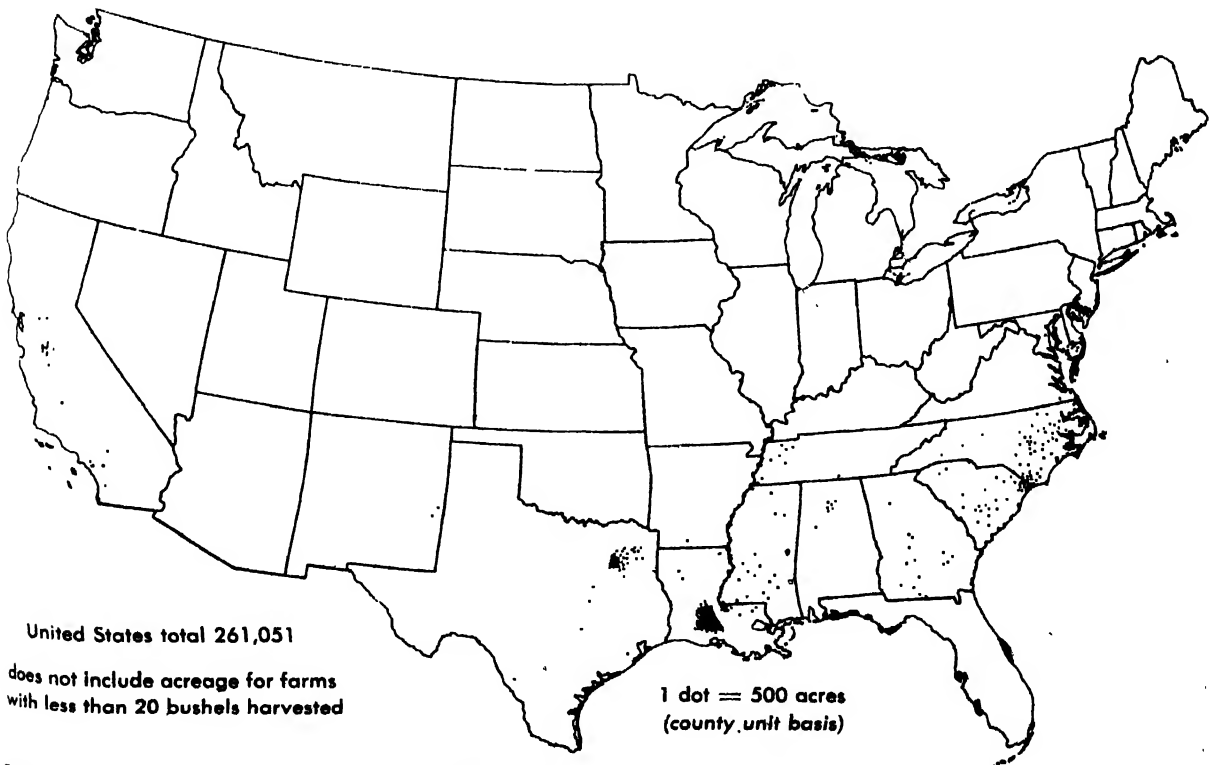


Fig. 2. Sweet-potato acreage in the United States for one year. (Bureau of the Census, U.S. Department of Commerce)

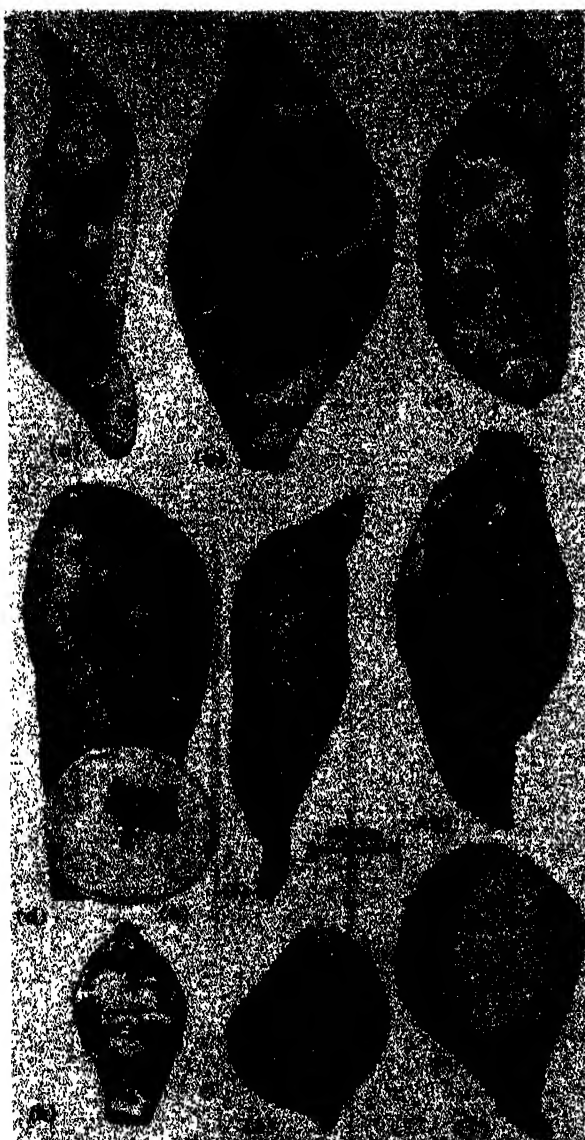


Fig. 3. Some important sweet-potato diseases. (a) Black rot. (b) Soft rot. (c) Soil rot. (d) Internal cork. (e) Section through internal cork-affected root. (f) Root knot. (g) Java black rot. (h) Scurf. (i) Circular spot. (j) Charcoal rot. (Louisiana Agricultural Experiment Station)

product that is similar to the Irish-potato chip, but higher in vitamins and carbohydrates. The most recent development is the making of sweet-potato flakes, similar to corn flakes, which can be used as cereal, in pie filling, in doughnuts, and in casserole form. In addition, many livestock growers use fresh and dehydrated sweet potatoes for feeding swine, poultry, and dairy cattle. The sweet potato stimulates milk production and increases the vitamin A content of the milk.

**True yam.** The true yam, *Dioscorea*, includes both edible and medicinal varieties. The latter are grown primarily for their cortisone, a steroid. A large number of cortisone yams are grown in the United States, particularly in Louisiana. See BREEDING (PLANT); CAROTENOID; STEROID; VEGETABLE GROWING; VITAMIN A; see also POTATO, IRISH.

[J.C.M.]

**Diseases.** Sweet-potato diseases, caused by fungi, nematodes, and viruses, affect both the plants growing in the field and the edible roots in storage and transit. Diseases incited by the different fungi are black rot, stem rot, soft rot, soil rot, scurf, circular spot, Java black rot, charcoal rot, foot rot, root rot, mottle necrosis, leaf blight, leaf spot, white rust, and true rust. The primary nematode disease is root knot. Virus diseases, important in sweet-potato production mainly since 1940, are mosaic, feathery mottle, internal cork, mottle leaf, dwarf, and others still unidentified.

Disease control, of any great importance in successful production and distribution of sweet potatoes, is difficult because sweet potatoes are propagated vegetatively. Essential control practices are selection of disease-free roots for propagation, chemical treatment of selected roots to destroy unseen traces of fungi, and crop rotation to reduce soil infestation. Soil treatments with fungicides or nematocides sometimes are necessary to rid infested soils of disease-producing organisms. Since 1940 resistant varieties obtained through breeding and selection have aided in control of stem rot, and to some extent soil rot, black rot, and root knot. World-wide search for disease-resistant strains among the numerous sweet-potato types available, combined with breeding programs, should result in development of more resistant varieties. See NEMATODA; PLANT DISEASE; PLANT VIRUS; see also FUNGISTAT AND FUNGICIDE. [W.J.M.]

## Potential, electric

At a given point in an electric field, the electric potential is the potential difference between that point and a place that is arbitrarily said to be at zero potential (see POTENTIAL DIFFERENCE). Frequently, it is convenient to consider that the earth is at zero potential, and this choice is made when convenience is served, as it usually is in a circuit analysis. In other cases, particularly in electrostatic fields, the problem is simplified if the potential is taken to be zero at a place infinitely far removed from the charges which produce the electric field. Using the latter choice, the potential  $V$  at a point  $P$  is the work per unit charge required to move a positive test charge from infinity to  $P$  (a test charge is one whose magnitude is small enough so that its presence does not distort the field being studied). Since electric field intensity  $E$  is force per unit charge, it follows that  $V$  at point  $P$  is given by the line integral

$$V = \int_{(\infty)}^{(P)} \mathbf{E} \cdot d\mathbf{s} = \int_{(\infty)}^{(P)} E \cos \theta \, ds \quad (1)$$

where  $d\mathbf{s}$  is a vector element of path length directed along the chosen path from  $\infty$  toward  $P$  and  $\theta$  is the angle between the vectors  $E$  and  $d\mathbf{s}$ . See ELECTRIC FIELD.

Potential is a scalar point quantity, since it has a magnitude only at every point in an electrostatic field, so a potential function for a particular problem is a scalar equation which expresses  $V$  as a

function of the coordinates in the electrostatic field (see ELECTROSTATICS).

**Principle of superposition.** This states that for any configuration of charges, the potential at a point in an electrostatic field is the algebraic sum of the potentials that each charge alone would produce at the point. Since this calls for an algebraic sum (rather than the vector sum required for the calculation of  $\mathbf{E}$ ), it is often easier to set up the potential function for a particular physical problem than to construct directly the function for  $\mathbf{E}$ . Then, with the potential function known, the differential relationship  $\mathbf{E} = -\text{grad } V$  is the one to use for calculation of  $\mathbf{E}$  rather than the integral relationship of Eq. (1). If  $V$  is expressed in Cartesian coordinates, this gradient equation is

$$\mathbf{E} = - \left( \mathbf{i} \frac{\partial V}{\partial x} + \mathbf{j} \frac{\partial V}{\partial y} + \mathbf{k} \frac{\partial V}{\partial z} \right) \quad (2)$$

where  $\mathbf{i}$ ,  $\mathbf{j}$ , and  $\mathbf{k}$  are unit vectors along the  $x$ ,  $y$ , and  $z$  axes, respectively.

**Equipotential surface.** As the name implies, this is an imaginary surface so drawn in an electric field that all points on it are at the same potential. Thus no electrical work is done in moving a charge from place to place on an equipotential surface. Electric lines of force are everywhere at right angles to equipotential surfaces. [R.P.WI.]

**Bibliography:** R. P. Winch, *Electricity and Magnetism*, 1955.

## Potential barrier

A field of force which surrounds the atomic nucleus and tends to keep bombarding particles out of the nucleus. If the bombarding particle is positively charged,  $+Ze$ , it will feel the Coulomb electrostatic repulsion to which corresponds a potential energy which varies as  $1/r$ , where  $r$  is the distance from the center of the nucleus,  $Z$  is the atomic number of the bombarding particle, and  $e$  is the charge of the proton. The potential barrier increases to the edge of the nucleus and is then overcome by the attractive nuclear forces. The maximum height of this Coulomb barrier is at the nuclear surface where it is

$$\frac{ZZ'e^2}{R}$$

$R$  being the radius of the nucleus and  $Z$  its charge. For protons, the barrier is about 4 Mev for neon and 17 Mev for uranium. If the initial kinetic energy of the incident particle is less than the barrier height, it can enter the nucleus only by virtue of what is called the quantum-mechanical tunnel effect.

There is an additional potential barrier called the centrifugal barrier for both charged and neutral particles if they have an orbital angular momentum  $lh/2\pi$  relative to the nucleus ( $l$  is the angular momentum quantum number and  $h$  is Planck's constant). This centrifugal barrier represents work done against the centrifugal force. It is propor-

tional to  $l(l+1)$  and varies as  $1/r^2$ ; thus it may be said to be thinner than the Coulomb barrier. See NUCLEAR STRUCTURE; SCHOTTKY EFFECT. [D.H.W.]

## Potential difference

The potential difference  $V_A - V_B$  between two points  $A$  and  $B$  in an electric field is defined as the change in potential energy of a test charge when it is moved between the two points, divided by the magnitude and sign of the test charge. A test charge is one whose magnitude is small enough so that its presence does not influence the field. Since electric field intensity  $E$  is force per unit charge and potential difference  $V_A - V_B$  is work per unit charge,  $E$  and  $V_A - V_B$  are related to each other by the line integral

$$V_A - V_B = - \int_{(B)}^{(A)} E \cos \theta \, ds$$

where  $ds$  is an element of path length from  $B$  toward  $A$ , and  $\theta$  is the angle between the vectors  $\mathbf{E}$  and  $d\mathbf{s}$ . See ELECTRIC FIELD.

The concept of potential difference is an important one in electric circuit calculations, where one usually defines it by saying that the potential difference  $V_A - V_B$  between two points  $A$  and  $B$  in the circuit is the work a unit charge will do in flowing from  $A$  to  $B$ . In a circuit, a voltmeter will give a direct measure of potential difference. See POTENTIAL, ELECTRIC. [R.P.WI.]

## Potentials (mathematics)

Suppose that situated at each point of a region  $R$  in space there is a vector, that is, a quantity which has both magnitude and direction. It is then said that there is a vector field in the region  $R$ . An example is given by the vector field consisting of the wind velocity and direction at each point in a region  $R$  near the surface of the earth. A vector field is represented in a rectangular coordinate system by an expression of the form  $a(x,y,z)\mathbf{i} + b(x,y,z)\mathbf{j} + c(x,y,z)\mathbf{k}$  where  $\mathbf{i}$ ,  $\mathbf{j}$ ,  $\mathbf{k}$  are unit vectors in the direction of the  $x$ ,  $y$ ,  $z$  axes, respectively. Suppose there is a single function  $\phi(x,y,z)$  such that  $\partial\phi/\partial x = a$ ,  $\partial\phi/\partial y = b$ ,  $\partial\phi/\partial z = c$ . Then the vector field is described completely by the single function  $\phi$  and in this case the field is called the gradient of  $\phi$ . Not every vector field is a gradient field since from the fact that  $\partial^2\phi/\partial x \partial y = \partial^2\phi/\partial y \partial x$  a necessary condition is  $\partial a/\partial y = \partial b/\partial x$ . Similar conditions relating  $a$  and  $c$  and relating  $b$  and  $c$  must also hold.

**Gravitation.** Newton's law of gravitation gives rise to a vector field in the following way. Imagine a body  $M_1$  of unit mass situated at the origin  $O$  of a rectangular coordinate system and another body  $M_2$  of unit mass situated at a point  $P(x,y,z)$  of space. According to Newton's law there is a force of attraction between  $M_1$  and  $M_2$  which is inversely proportional to the square of the distance  $r = \sqrt{x^2 + y^2 + z^2}$  between the bodies. For convenience the units are selected so that the magnitude of the force is equal to  $1/r^2$ . The components of

the force at  $P$  are

$$-\frac{1}{r^2} \cos \alpha, -\frac{1}{r^2} \cos \beta, -\frac{1}{r^2} \cos \gamma$$

where  $\alpha, \beta, \gamma$  are the angles that the line  $OP$  makes with the coordinate axes. These may be written, also, as

$$-\frac{x}{r^3} \mathbf{i} - \frac{y}{r^3} \mathbf{j} - \frac{z}{r^3} \mathbf{k}$$

As the body  $M_2$  moves throughout space a vector field or force field is generated. This is usually designated the Newtonian or gravitational field. A simple calculation shows that the function  $\phi = 1/r$  has as its gradient the gravitational field. In other words by means of a single function the gravitational forces due to a unit mass at a single point are completely described. This function is called the unit gravitational potential or simply the potential.

In the idealized situation just described the masses  $M_1$  and  $M_2$  occupy single points in space. Let  $M_1$  occupy a domain  $D$  in space and let  $\rho(\xi, \eta, \zeta)$  be the density of  $M_1$  at each point  $(\xi, \eta, \zeta)$  of  $D$ . Once again  $M_2$  is a unit mass located at a point  $P(x, y, z)$  not in  $D$ . The mass  $M_1$  may be decomposed into elementary pieces and the gravitational potential at  $P$  due to  $M_1$  will be the sum of the potentials of the individual parts. By the usual limiting process of calculus the potential  $\phi$  is given by

$$\phi(x, y, z) = \iiint_D \frac{\rho(\xi, \eta, \zeta) d\xi d\eta d\zeta}{[(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2]^{1/2}} \quad (1)$$

The gradient of  $\phi(x, y, z)$  yields the gravitational force field due to the mass  $M_1$  at each point of space exterior to  $D$ . Although the integrand in (1) becomes singular for points  $P(x, y, z)$ , within  $D$  the integral itself nevertheless has a finite value. Thus it is possible to consider the potential of solid objects throughout the entire space.

A simple calculation shows that the function  $\phi = 1/r$  satisfies the Laplace equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 0$$

at all points except  $r = 0$ . Similarly the integral (1) satisfies the same equation at all points not in  $D$ . Thus the study of gravitational fields is intimately connected with the study of a particular partial differential equation—the Laplace equation. This equation arises in many branches of mathematical physics, and hence an immediate analogy is drawn between the Newtonian potential and potentials which arise in the study of electricity, magnetism, and fluid flow.

**Electricity.** The force field due to the attraction and repulsion of electric charges admits of an analysis similar to that for the gravitational potential. The main distinction between Coulomb's law and Newton's law lies in the fact that similarly charged

electric particles repel each other. The potential function for two particles each of unit charge is either  $1/r$  or  $-1/r$  according as they are of opposite or of like charge. If one considers the field due to a number of charges, some positive and some negative, care must be taken in regard to the signs in computing the electric potential.

**Magnetism.** The field of force due to a magnet is somewhat more complicated than those due to masses or electric particles. Under ordinary considerations of a magnet it is essential that there be both a north and a south pole. There is no parallel to a point unit mass or a positive electric charge of one unit. For this purpose it is necessary to introduce the idealized concept of a magnetic particle. Imagine one pole of a magnet located at a single point  $P_0(x_0, y_0, z_0)$  with an attracting strength  $m$  and the other end of the magnet at the point  $P_1(x_1, y_1, z_1)$  with attracting strength  $-m$ . The potential function at a point  $P(x, y, z)$  in space due to this magnet is given by the function

$$\phi = (x, y, z) = m/r_0 - m/r_1$$

where

$$r_0^2 = (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 \\ r_1^2 = (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2$$

In other words, idealized magnets also follow Newton's or Coulomb's laws with strength of poles replacing masses and charged particles. One can denote by  $d$  the length of the segment  $P_0P_1$ , let the point  $P_1$  approach  $P_0$  along the line segment  $P_0P_1$ , and let  $m$  approach infinity in such a way that  $m \cdot d$  is always equal to a constant  $\mu$ . The function  $\phi$  will approach a limit which is a potential function. The quantity  $\mu$  is called the moment of the magnetic particle at  $P_0$ . It is clear that  $\phi$  also depends on the direction along which  $P_1$  approaches  $P_0$  and this direction is called the axis of the particle. It is not difficult to see that the potential of a magnetic particle is the moment times the directional derivative of the function  $1/r$  in the direction of its axis. The potentials due to curves, surfaces, and solids made up of magnetic particles are obtained by integration as in the case of gravitational potentials.

**Heat conduction.** Heat conduction in a solid is governed by the partial differential equation

$$\frac{\partial \phi}{\partial t} = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2}$$

In the steady state flow of heat  $\partial \phi / \partial t = 0$ , and the solution, independent of time, satisfies the Laplace equation. Thus the problem of determining the steady state flow of heat in a body is equivalent to the problem of finding a potential function under appropriate boundary conditions.

**Fluid flow.** Suppose a fluid flows through a region  $R$  in space. Let  $u(x, y, z, t)$ ,  $v(x, y, z, t)$ ,  $w(x, y, z, t)$  be the components of the velocity vector at the point  $P(x, y, z)$  at time  $t$ . If the flow is stationary the functions  $u$ ,  $v$ ,  $w$  do not depend on  $t$ . If in addition the fluid is incompressible and the flow irro-

tational, that is, free of vortices, there will exist a potential function  $\phi$  such that its gradient is the velocity field which describes the flow.

**Other concepts.** Many problems of potential theory become simpler if a restriction to two dimensions is made. Suppose a wire of constant linear density  $P$  is situated along the  $z$  axis. One can assume that it is infinitely long, extending from  $z = -\infty$  to  $z = +\infty$ . The force field at a point  $P(x, y, 0)$  is calculated to be  $2\rho/r$  where  $r^2 = x^2 + y^2$  and is directed along the line  $OP$ . The potential which yields this force field is  $\phi(x, y) = 2\rho \ln(1/r)$  and this logarithmic potential plays the same role in two dimensions that the function  $1/r$  does in the three-dimensional case. In two variables the function  $\ln(1/r)$  satisfies the Laplace equation for  $r \neq 0$  while the function  $1/\sqrt{x^2 + y^2}$  does not.

An important concept in potential theory is the notion of equipotential surface. If  $\phi(x, y, z)$  is a potential function, then in general  $\phi(x, y, z) = C_0$ , where  $C_0$  is a constant, defines a surface in space. From analytic geometry it is known that if  $F(x, y, z) = 0$  is any surface the vector

$$\frac{\partial F}{\partial x} \mathbf{i} + \frac{\partial F}{\partial y} \mathbf{j} + \frac{\partial F}{\partial z} \mathbf{k}$$

represents a normal to the surface at the point  $(x, y, z)$ . In the case of a potential function this vector is simply

$$\frac{\partial \phi}{\partial x} \mathbf{i} + \frac{\partial \phi}{\partial y} \mathbf{j} + \frac{\partial \phi}{\partial z} \mathbf{k}$$

or the gradient field of the potential. This means that at every point of the surface  $\phi(x, y, z) = C_0$  the force or flux vector is perpendicular to the surface. These surfaces are called equipotential surfaces, and they are of great interest in problems of electricity, fluid flow, and heat conduction.

In order to exhibit the close relationship between solutions of the Laplace equation (called harmonic functions) and potentials a sketch will be given which shows that every harmonic function may be represented as a sum of potentials. Denote by  $\Delta$  the operator

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

If  $R$  is a region with smooth boundary  $S$  and  $n$  denotes the outer normal to  $S$ , the following identity is a consequence of Green's theorem.

$$\iiint_R (u \Delta v - v \Delta u) dV = \iint_S \left( u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) dS \quad (2)$$

for any two functions  $u$  and  $v$  which have continuous second derivatives in a region containing  $R$ . Let  $P_0(x_0, y_0, z_0)$  be a fixed point of  $R$  and  $r$  the distance from  $P(x, y, z)$  to  $P_0$ . The selection  $1/r$  is made for the function  $v$ . This introduces a certain difficulty in relation (2) above because  $v$  has a sin-

gularity at  $P_0$ . However (2) still holds if a small sphere about  $P_0$  is deleted from  $R$  and to the boundary  $S$  is added the surface of this sphere. After some reduction the resulting formula is

$$u(x_0, y_0, z_0) = -\frac{1}{4\pi} \iiint_R \frac{\Delta u}{r} dV + \frac{1}{4\pi} \iint_S \frac{\partial u}{\partial n} \frac{1}{r} dS - \frac{1}{4\pi} \iint_S u \frac{\partial}{\partial n} \left( \frac{1}{r} \right) dS$$

If now  $u$  is harmonic the first integral on the right vanishes and the relation

$$u(x_0, y_0, z_0) = \frac{1}{4\pi} \iint_S \frac{\partial u}{\partial n} \frac{1}{r} dS - \frac{1}{4\pi} \iint_S u \frac{\partial}{\partial n} \left( \frac{1}{r} \right) dS$$

follows. The first integral represents the Newtonian potential due to a surface  $S$  of density  $\partial u / \partial n$  while the second integral is the potential of a surface  $S$  of magnetic particles with magnetic moment  $u(x, y, z)$ . See CALCULUS OF VECTORS; DIFFERENTIAL EQUATION; LAPLACE'S DIFFERENTIAL EQUATION; OPERATOR THEORY; POTENTIALS (PHYSICS).

[M.H.P.]

*Bibliography:* O. D. Kellogg, *Foundations of Potential Theory*, 1929; H. Poincaré, *Théorie du Potentiel Newtonien*, 1899.

## Potentials (physics)

Potentials in physics are of two kinds: (1) measures of storage of a quantity available for possible use, and (2) functions whose rates of change are the primary quantities of a physical theory. Vague as the foregoing definition is, it is scarcely broad enough to include all current usage of the word potential. In the oldest and most common applications, both senses (1) and (2) apply, and in sense (2) the potential is a single function from which a vector, such as a force or a velocity, is derived. In the newer and broader usage,  $A$  is a potential of  $B$  if some combination of the partial derivatives of  $A$  equals  $B$ , irrespective of physical interpretation. In many cases aspect (1) leads to, or is replaced by, a conservation principle, asserting the existence of a quantity which remains unchanged. The conserved quantity in some cases is defined in terms of the potential and in some cases is not. Sometimes the existence of a potential leads also to some principle of economy, whereby some quantity defined by the aid of the potential is less, according to the particular physical theory, than it would be if that theory did not hold.

It cannot justly be said that the word potential stands for any physical concept. Rather, as now used, it refers loosely to the mathematical simplification of formulas that sometimes results when some of the differential equations of a theory are replaced by their general solutions in terms of arbitrary functions. These functions are then called potentials.

Absence of a unifying concept makes an enumeration of typical cases the only sensible way to describe the subject.



**Potentials in mass-point mechanics.** A body of constant mass  $M$  near the earth's surface obeys the law of conservation of energy

$$\text{Total energy} = T + U = \text{const} \quad (1)$$

Here  $T$  is the kinetic energy, arising only from the motion of the body:  $T = \frac{1}{2} Mv^2$ , where  $v$  is the speed. The energy  $U$  is a potential energy, arising only from the location of the body:  $U = Mgz$ , where  $z$  is the height above some arbitrary level and  $g$  is the constant acceleration of gravity. Loss of some potential energy results in gain of an equal amount of kinetic energy, and vice versa. The force  $F$  which acts on the body is given by  $F = -\text{grad } U = -\partial U / \partial r$ , where  $r$  is the position vector. Both aspect (1) and aspect (2) of the definition of potential are illustrated here.

More generally, consider a system of  $n$  mass-points with masses  $M_k$  and position vectors  $r_k$ . A function  $U(r_1, r_2, \dots, r_n)$  is said to be a potential energy of the system if the force acting upon the  $k$ th mass point is given by

$$F_k = -\text{grad}_k U = -\frac{\partial U}{\partial r_k} \quad (2)$$

Therefore, the equations of motion are

$$M_k \ddot{r}_k = -\frac{\partial U}{\partial r_k} \quad k = 1, 2, \dots, n \quad (3)$$

Integration of Eqs. (3) again yields the theorem of conservation of total energy in the form of Eq. (1), where

$$T = \frac{1}{2} \sum_{k=1}^n M_k \dot{r}_k^2$$

This result illustrates not only aspects (1) and (2) but also the simplicity of specifying the motion of the system by means of the single function  $U$  instead of the  $n$  vectors  $F_k$ . See ENERGY.

One principle which may serve as the basis of statics states that the equilibrium position of a conservative system is such that its potential energy assumes the least value possible. There are several generalizations of this idea to dynamics. For example, let the conservative system considered be described by  $3n$  generalized coordinates  $q_k$ . The kinetic potential  $L(q_1, \dots, q_{3n}, \dot{q}_1, \dots, \dot{q}_{3n})$  is defined by  $L = T - U$ . The motion of the system is such as to give the total kinetic potential its least possible value. A mathematical expression for this requirement is

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k} - \frac{\partial L}{\partial q_k} = 0 \quad k = 1, 2, \dots, 3n \quad (4)$$

See LAGRANGE'S EQUATIONS.

**Gravitational potentials.** One expression for the classical law of universal gravitation is that there exists a potential energy  $U(P)$  of the form

$$U(P) = G \int \frac{dm}{r(dm, P)} \quad (5)$$

where  $r(dm, P)$  is the distance from the element of mass  $dm$  to the point  $P$ . Here  $G$  is the constant of universal gravitation. The integration is carried out over all space; both discrete and continuous masses may be included. When all masses are discrete, the theory becomes a special case of that expressed by Eqs. (3). When mass is smoothly distributed in space, so that  $dm = \rho dv$ , where  $\rho$  is the density and  $dv$  is an element of volume, it can be shown from Eq. (5) that

$$\nabla^2 U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 4\pi G\rho \quad (6)$$

In empty space,

$$\nabla^2 U = 0 \quad (7)$$

The force acting on a unit mass present in, but not influencing, the gravitational field is  $-\text{grad } U$ . The extensive mathematical developments based on Eqs. (6) and (7) constitute potential theory. See GRAVITATION.

In general relativity, the metric tensor of space-time is supposed to be determined by the distribution of energy and momentum. Its ten components  $g_{km}$  are called gravitational potentials. These quantities do not conform to the definition directly; rather, their name arises from the fact that it is possible to define from them, mathematically, a quantity which reduces, in the classical approximation, to the classical potential  $U$  which satisfies Eq. (6). See RELATIVITY.

**Fluid potentials.** If the motion of a fluid or other deformable substance is such that there exists a function  $\phi$  giving the velocity  $v$  as follows:

$$v = -\text{grad } \phi \quad (8)$$

then  $\phi$  is said to be a velocity potential. A special kind of motion, called irrotational motion, is characterized by Eq. (8). In such a motion, the small portions of the material are not spinning, though they may circulate. If the substance is incompressible, then  $\nabla^2 \phi = 0$ ; compare Equation (7). In this case, the motion has less kinetic energy than any other one corresponding to the same discharge of matter through a fixed bounding surface. See LAPLACE'S IRROTATIONAL MOTION.

More generally, if the acceleration  $a$  satisfies a relation such as Eq. (8),  $a = -\text{grad } \Phi$ , then  $\Phi$  is called an acceleration potential. Again a special kind of motion is characterized; this time, the circulation of velocity around any closed ring of particles remains unaltered. This is a conservation principle.

Virtually all of the science of hydrodynamics and aerodynamics concerns motions of these kinds. The acceleration potential may generalize the potential energy, since in some conditions  $\Phi = Y + \int dp/\rho$ , where  $Y$  is a potential energy and  $p$  is the pressure. In a steady motion, along each streamline

$$\frac{1}{2} v^2 + \Phi = \text{const} \quad (9)$$

This is a theorem of conservation of energy somewhat like Eq. (1). For a motion having a velocity potential, the constant in Eq. (9) has the same value for each streamline. See STREAMLINE FLOW.

**Elastic and plastic potentials.** In a perfectly elastic body the stress, or distribution of interior forces, arises solely in response to the deformation undergone from the free or natural state. The work done in producing the deformation is available for reversing it; this work is the stored energy. If suitable measures of stress  $T$  and strain  $E$  are selected, then the stored energy  $F(E)$  is also a potential for the stress, as is expressed by the symbolic formula

$$T = \frac{\partial F}{\partial E} \quad (10)$$

A similar formula holds in some theories of plasticity, except that  $F$  depends upon the velocity of deformation rather than upon the deformation itself; such a function is called a plastic potential.

**Thermodynamic potentials.** The internal energy  $\epsilon$  of thermodynamics is related to the stored energy of elasticity theory. While a rather general treatment is possible, only the case of a simple fluid is presented here. The internal energy is then determined by the specific volume,  $v$ , and the specific entropy,  $\eta$ , so that  $\epsilon = \epsilon(v, \eta)$ . The temperature  $\theta$  and the pressure  $\pi$  are then obtained from  $\epsilon$  as a potential:

$$\theta = \frac{\partial \epsilon}{\partial \eta} \quad \pi = -\frac{\partial \epsilon}{\partial v} \quad (11)$$

Many other thermodynamic potentials may be introduced. The simplest are the free energy,  $\psi$ , the enthalpy,  $\chi$ , and the free enthalpy,  $\zeta$ , defined as follows:

$$\begin{aligned} \psi &= \epsilon - \eta\theta \\ \chi &= \epsilon + \pi v \\ \zeta &= \epsilon + \pi v - \eta\theta \end{aligned} \quad (12)$$

Therefore  $\epsilon - \psi - \chi + \zeta = 0$ . The functions  $\epsilon(v, \eta)$ ,  $\psi(\theta, v)$ ,  $\chi(\eta, \pi)$ , and  $\zeta(\theta, \pi)$  are thermodynamic potentials in the sense that all thermodynamic functions may be obtained as combinations of the derivatives of any one of them. See THERMODYNAMIC PROCESSES.

**Electromagnetic potentials.** Since the law of electrostatic force between charge-points is of the same mathematical form as that between mass-points, there exists an electrostatic potential of the same form as the function  $U$  given by Eq. (5), except that the element of mass  $dm$  is replaced by an element of charge  $de$  which may assume negative as well as positive values, and the constant  $G$  is replaced by an electrostatic one. Thus the whole theory of Newtonian potentials is applicable, by mere changes of wording, to electrostatics.

But the subject may be approached from a more profound viewpoint. Two of the four vectorial equations governing the electromagnetic field may

be written in the forms

$$\text{div } B = 0 \quad \text{curl } E = -\frac{\partial B}{\partial t} \quad (13)$$

where  $B$  is the magnetic induction and  $E$  is the electric field intensity. The first of these conditions, regarded as a partial differential equation, has the general solution

$$B = \text{curl } A \quad (14)$$

where the arbitrary function  $A$  is called the vector potential. Substituting Eq. (14) into the second of Eq. (13) yields

$$\text{curl} \left( E + \frac{\partial A}{\partial t} \right) = 0 \quad (15)$$

The general solution of this partial differential equation is

$$E + \frac{\partial A}{\partial t} = -\text{grad } \Phi \quad (16)$$

Thus the electrostatic potential is a special case of the general electrodynamic scalar potential  $\Phi$ .

The potentials  $\Phi$  and  $A$  have to be replaced by suitable tensor potentials when electrodynamics is expressed in a four-dimensionally invariant formalism. See MAXWELL'S EQUATIONS.

**Potentials of arbitrary fields.** There are various representations of an arbitrary vector field  $c$ . Some of these are

$$\begin{aligned} c &= \text{grad } H + F \text{ grad } G \\ &= \text{grad } M + \text{grad } K \times \text{grad } L \\ &= -\text{grad } S + \text{curl } P \end{aligned} \quad (17)$$

The existence of the various potentials denoted by capital letters in these formulas reflects geometric properties of vector fields in general and finds application in particular theories according to how those geometric properties are given physical interpretation. There are similar potentials for tensors.

Although formulas (17) are written for three-dimensional space, similar potential representations are valid in spaces of any dimension, and in particular, for space-time.

For example, if  $\delta$  is a vector density and  $\phi$  is a bivector satisfying the conservation principles

$$\text{Div } \delta = 0 \quad \text{Curl } \phi = 0 \quad (18)$$

respectively, then there exist potentials  $\beta$  and  $\alpha$  such that

$$\delta = \text{Div } \beta \quad \phi = \text{Rot } \alpha \quad (19)$$

where Div, Curl, and Rot are appropriately defined differential operators. Rot, for example, is defined in the work by Schouten listed in the bibliography.

**Retarded potentials.** A special kind of four-dimensional potential appears in the many theories that lead to the equation for linear waves,

$$\frac{1}{c^2} \frac{\partial^2 W}{\partial t^2} = \nabla^2 W \quad (20)$$

$c$  being the speed at which disturbances such as waves of sound or light travel (see WAVE EQUATION; WAVE MOTION). Then

$$W(P, t) = \frac{1}{4\pi} \oint_S \left[ \frac{1}{r} \frac{\partial W^*}{\partial n} - \left( W^* + \frac{r}{c} \frac{\partial W^*}{\partial t} \right) \frac{\partial}{\partial n} \left( \frac{1}{r} \right) \right] dA \quad (21)$$

where  $S$  is a closed surface,  $r$  is the distance from the point  $P$  to  $dA$ , and the asterisk indicates that the time  $t$  is to be replaced by the earlier time  $t - r/c$ . This formula, which shows precisely how signals traveling at the speed  $c$  propagate to the point  $P$  conditions formerly holding upon the surrounding surface  $S$ , is said to represent the wave function  $W$  in terms of retarded potentials. See CALCULUS OF TENSORS; CALCULUS OF VECTORS; POTENTIALS (MATHEMATICS).

**Bibliography:** W. V. Houston, *Principles of Mathematical Physics*, 2d ed., 1948; O. D. Kellogg, *Foundations of Potential Theory*, reprint, 1954; H. B. Phillips, *Vector Analysis*, 1933; J. A. Schouten, *Ricci-Calculus*, 2d ed., 1954.

## Potentiometer (variable resistor)

A variable resistance device with three terminals used in electric circuits. As shown schematically in Fig. 1, the three terminals are the two ends of a resistor (or series combination of resistors) and a movable connection, which allows adjustment of the resistance between this movable connection and either end connection. The movable connection often consists of a sliding contact which moves along the actual resistor element. The size or rating of a potentiometer is specified by giving its total resistance in ohms and the permissible losses in watts (see RESISTOR). By using only the movable and one fixed connection, a potentiometer may be used as a rheostat. See RHEOSTAT.

The term potentiometer is also applied to a precision instrument used to measure or compare electrical voltages; for a discussion of this device which depends on the same type of resistor arrangement, see POTENTIOMETER (VOLTAGE MEASUREMENT).

**Use.** A potentiometer is used to adjust and control the electric potential difference (voltage) applied to some device or part of a circuit. The output voltage may be varied from zero to the value of the input voltage. Examples of its use are as a field-current control on an electric generator and as a

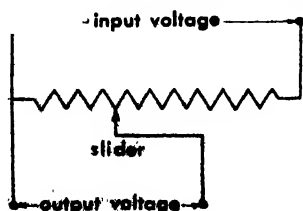


Fig. 1. Potentiometer schematic.

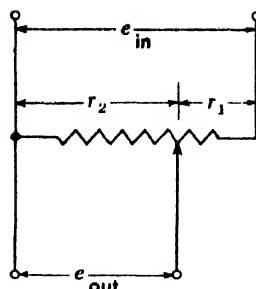


Fig. 2. Voltage divider.

volume control on a radio. Since the resistance between the input terminals is fixed (assuming the load device takes little current), potentiometers with precision resistors are used as range selector switches on vacuum-tube voltmeters and other precision electronic measuring equipment. Other uses are to divide a voltage into two parts, to compare two voltages, and to divide the total resistance between two parts of a circuit. The ratio between output and input voltage, as shown in Fig. 2, is

$$\frac{e_{out}}{e_{in}} = \frac{r_2}{r_1 + r_2}$$

**Construction.** A potentiometer may be linear or nonlinear. In the linear case the resistor is uniform, and the voltage distribution along the resistor is the same for any fixed fraction of its total length. Therefore, the output voltage (Fig. 1) is proportional to the slider position. In a nonlinear potentiometer the resistance per unit length varies, and the output voltage varies as some function (such as the logarithm, the square, or the sine) of the slider position. Nonlinear potentiometers are often called tapered potentiometers. In some cases the current-carrying capacity of various parts of the potentiometer may be different.

A slide-wire potentiometer employs a movable sliding connection on a length of resistance wire.

A wire-wound potentiometer is similar to a slide-wire one, except that the resistance wire is wound on a form and contact is made by a slider which moves along an edge from turn to turn. The form may be straight or bent into a part of a circle, in which case the slider is mounted on an arm which is rotated by a knob. The form may be made of a ceramic material for heat resistance or a good grade of stiff paper or plastic (known as a card).

A carbon potentiometer uses a thin layer of carbon or graphite in place of a resistance wire.

A button-type potentiometer uses fixed contact points which are touched by the slider. Fixed resistors are then connected between the buttons.

**Multiturn potentiometers.** Sold under various trade names, these have the resistance material, usually coiled resistance wire, placed in the form of a cylindrical helix. A slider is moved down the helix by a lead-screw arrangement. This arrangement permits a long length of potentiometer in a small volume and gives more accurate adjustment than is possible with a single-turn potentiometer.

**Trimmer potentiometer.** This is a potentiometer used to provide a small percentage adjustment and is often used with a coarse control. [C.F.]

## Potentiometer (voltage measurement)

A device for the measurement of an electromotive force (emf) by comparison with a known potential difference (see ELECTRICAL MEASUREMENTS; POTENTIAL, ELECTRIC). The known potential difference is established by the flow of a definite current through a known resistance, using a standard cell as a reference. The principal potentiometer circuits are (1) the constant-current dc potentiometer (historically known as Poggendorff's first method); (2) the Brooks deflectional dc potentiometer (a variant of the basic constant-current potentiometer); (3) the constant-resistance dc potentiometer (known as Poggendorff's second method); (4) the Drysdale ac potentiometer; and (5) the Tinsley-Gall ac potentiometer.

**Constant-current dc potentiometer.** This is widely used in the standardization of dc measuring instruments; its basic circuit is illustrated in Fig. 1.

A battery causes a current  $I_{ab}$  to flow through a calibrated resistor or slidewire  $R_{ab}$ . With the switch  $S$  connected to the standard cell and sliders  $a'$  and  $b'$  set for an appropriate value of resistance  $R_1$ , the slidewire current  $I_{ab}$  is adjusted until the galvanometer current  $I_g$  is equal to zero. The slidewire voltage  $R_1 I_{ab}$  is now equal to the standard cell emf  $E_s$ .

The unknown voltage  $V$  is then substituted for the standard cell and the sliders  $a'$  and  $b'$  are again adjusted for a null galvanometer reading, with slidewire current  $I_{ab}$  unchanged. Let the new value of slidewire resistance be  $R_2$ . The unknown voltage is

$$V = E_s R_2 / R_1 \quad (1)$$

Thus, this potentiometer compares potential differences in terms of known resistances, which are usually calibrated in terms of volts and millivolts.

A complete potentiometer circuit is shown in Fig. 2. The resistance of the dial switch, the main slidewire, and the standard slidewire represent  $R_{ab}$ . Transfer from the standardizing to the measuring circuit is accomplished by the switch  $U$ . With the switch in the left position, the standard cell, with voltage  $E_s$ , and the galvanometer are connected across the resistance composed of nine coils and part of the slidewire. Balance is obtained by adjustment of the regulating rheostats. The switch

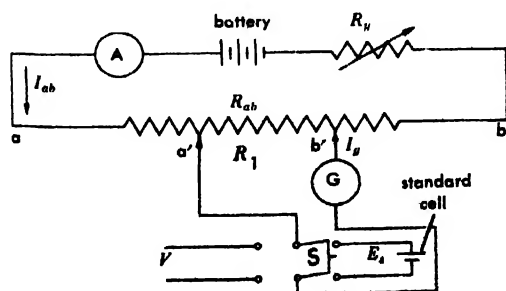


Fig. 1. Elementary dc potentiometer circuit. (From I. F. Kinnard, *Applied Electrical Measurements*, Wiley, 1956)

is then thrown to the right, connecting the unknown voltage  $E_x$  and the galvanometer across the dial switch and main slidewire, both of which are adjusted to obtain balance. A range switch reduces the slidewire current to  $\frac{1}{10}$  or  $\frac{1}{100}$  of its full value when required. The measurement range of the constant-current potentiometer is 0–2.0 volts. Multipliers of 1, 0.1 and 0.01 are available. This device has undergone intensive refinement, and errors can be reduced to the order of 0.025% of the reading. Potentiometer measurement of current is accomplished by passing current through a standardized resistor of appropriate value and measuring the potential difference across this resistor.

**Brooks deflectional dc potentiometer.** This potentiometer eliminates the time-consuming operation of obtaining an exact balance of the slidewire. This is accomplished by circuitry and a galvanometer calibration by which the galvanometer reading may be added to or subtracted from an approximate dial setting. As illustrated in Fig. 3,  $E_x$  is the unknown voltage,  $E_s$  is the terminal voltage of a storage cell of negligible resistance, and  $R_g$  is the resistance of galvanometer plus its series resistor. Analysis of the circuit shows that if the quantity  $R_g + R_1(R_2 + R_n)/R_1 + R_2 + R_n$  is kept constant for all positions of the sliders, the galvanometer can be calibrated directly in volts. This is achieved by the addition of the auxiliary resistor  $R'_g$  in series with the galvanometer and slider. As the slider is moved, an increment of resistance  $\Delta R_g$  is added to or subtracted from the above resistance to keep the ratio constant. The total unknown emf is the algebraic sum of the slider setting and galvanometer reading. The instrument is otherwise similar to the basic constant-current potentiometer. Self-contained deflectional potentiometers are available in ranges of 0–1.5 volts, which may be extended to 300 volts, and 0–0.75 amp, which may be extended to 150 amp with special multipliers. Maximum errors of 0.05% self-contained, plus 0.04% for multipliers, if used, are obtained.

**Constant-resistance dc potentiometer.** This instrument is especially adapted to the measurement of very low potentials. As illustrated in Fig. 4, the unknown potential  $E_x$  is applied to the known constant resistance  $R_2$ , and the current  $I$  is adjusted for null reading of the galvanometer. In series with  $R_2$  is another constant and known resistance  $R_1$ . The potential  $E_p$  across this resistance is measured by any null-type potentiometer. Then

$$E_x = \frac{E_p R_2}{R_1} \quad (2)$$

Thus, the range of the null-type potentiometer may be extended downward many fold by making  $R_1$  large and  $R_2$  small. Reliable measurement in the microvolt range is practicable if care is taken to avoid parasitic emfs in  $R_2$ . The null potentiometer may be replaced by a dc voltmeter, but measurement accuracy is thereby diminished.

**Drysdale ac polar potentiometer.** This is shown in principle in Fig. 5. An ordinary slidewire is supplied with current by a phase-shifting trans-

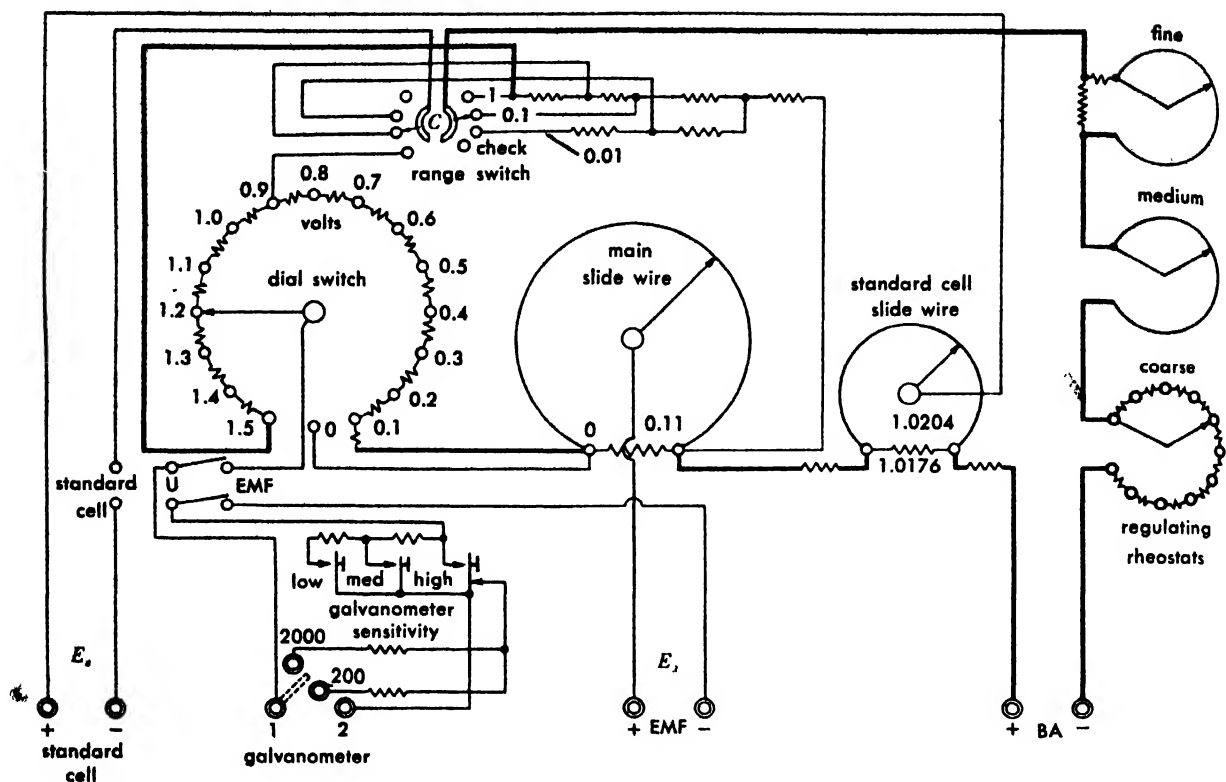


Fig. 2. Complete dc potentiometer circuit. (Leeds and Northrup Co.)

The amount of current is measured by an ammeter which must be an accurate ac to dc transfer instrument. The potentiometer current and voltage drop along the slidewire are brought into phase with the unknown voltage by adjustment of the transformer rotor. The unknown voltage is then measured by observation of the slidewire setting for a null indication of the vibration galvanometer. The potentiometer is calibrated on direct current by use of a dc potentiometer. The equivalent value of ac, as measured by ammeter A, is maintained for ac measurement. Maximum errors of 0.1–0.2% are obtained, depending upon the accuracy of the ammeter and sensitivity of the galvanometer.

**Tinsley-Gall ac polar potentiometer.** Shown in Fig. 6, this potentiometer measures the unknown emf by measurement of the in-phase and quadrature components in reference to a standard current. Two potentiometers are used, the currents in which are numerically equal but 90° displaced in phase. The in-phase potentiometer is first standardized

by closing switch  $S_1$  to the left and adjusting  $R_1$  for null reading of galvanometer G. Then

$$E_s = I_1 R_g \quad (3)$$

$R_g$  is preadjusted so that a definite current, usually 50 milliamperes (ma), flows through the slidewire. The reading of the reflecting dynamometer is then observed. Alternating current from the in-phase source is then applied by closure of switch  $S_1$  to the right, and rheostat  $R_2$  is adjusted to reproduce the 50-ma reading of the reflecting dynamometer. The quadrature potentiometer is now standardized by closure of  $S_2$  in position 1, together with closure of contacts ab and cd. This series connects the quadrature slidewire through the vibration galvanometer and the in-phase slidewire to the secondary of mutual inductor  $M$  in which an emf

$$E_m = j\omega M I_2 \quad (4)$$

is generated. Reversing switches  $S_3$  and  $S_4$  are connected so that the voltage drops  $e_1$  and  $e_2$  are opposed. The in-phase slidewire is set to a predetermined value of induced voltage  $E_m$  as determined by the frequency and coefficient of mutual induction. The rheostat  $R_3$  is then adjusted for null indication of G, and 50 ma now flow through the quadrature slidewire. The transfer switch  $S_2$  is now changed to position 2, cutting out the mutual inductor and series connecting the slidewire switches  $S_2$  and  $S_3$ , as well as the vibration galvanometer to the unknown voltage  $V_u$ . The apparatus is now ready for measurement of the quadrature components of the unknown voltage. All four quadrants may be explored by use of reversing

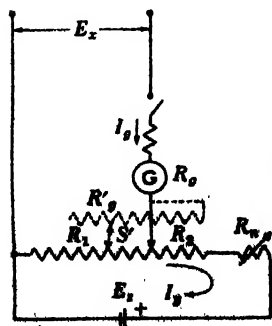


Fig. 3. Elementary Brooks deflection potentiometer circuit. (From I. F. Kinnard, *Applied Electrical Measurements*, Wiley, 1956)

and  $S_4$ . This potentiometer, for best results, requires a very stable ac power supply consisting of two single-phase alternators with adjustable stators. It is used especially for such specialized laboratory procedures as instrument-transformer ratio, phase-angle determination, and ac cable testing.

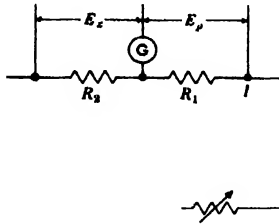


Fig. 4. Elementary constant-resistance potentiometer circuit. (From I. F. Kinnard, *Applied Electrical Measurements*, Wiley, 1956)

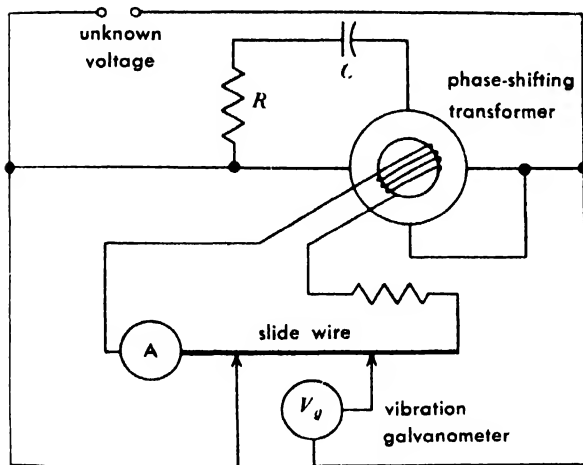


Fig. 5. Drysdale ac potentiometer circuit. (From I. F. Kinnard, *Applied Electrical Measurements*, Wiley, 1956)

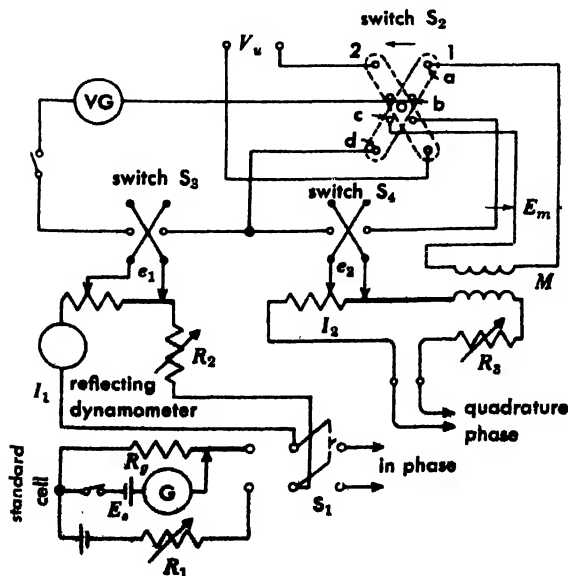


Fig. 6. Tinsley-Gall ac potentiometer circuit. (From I. F. Kinnard, *Applied Electrical Measurements*, Wiley, 1956)

For other methods of voltage measurement, see **VOLTMETER**. [A.J.C.]

**Bibliography:** I. F. Kinnard, *Applied Electrical Measurements*, 1956.

## Pottery

Vessels made entirely or partly of clay, and fired to a strong, hard product; occasionally, the term refers to just the lower grades of such ware. Alternatively, it refers to the manufacturing plant at which such ware is made. An older meaning is the art of making such ware; in this use it becomes synonymous with the older definition of ceramics (see **CERAMIC TECHNOLOGY**). Pottery may be glazed or unglazed (see **GLAZING**).

Grades of pottery are distinguished by their color, strength, absorption (the weight of water soaked up when the piece is submerged, expressed as a percentage of the original weight), and translucency (ability to pass light). All these properties refer to the material or "body" under any glaze present. See **PORCELAIN**.

China is white in color, strong, has less than 2% absorption, is always glazed, and is translucent in thin ware. Special types are bone china, containing phosphates from calcined bones as a fluxing material; hotel china, made extra-thick for maximum strength and therefore not translucent; frit china (also called frit porcelain), containing ground glass to give a translucent body maturing at a moderate firing temperature.

Stoneware has a cream or brown color, high strength, 0-5% absorption, and no translucency; it is often unglazed. Chemical ware is an example.

Earthenware (sometimes known as semivitreous china) is white or ivory, has less strength than china or porcelain, 3-10% absorption, and no translucency; it is usually glazed. Most everyday tableware is of this type. Faience is a special type with a soft, porous, red or yellow body covered by an opaque glaze. Majolica ware is a type having over 15% absorption and an opaque glaze over a relatively weak red or gray body.

Other special types of pottery are Parian ware, a body with a high flux content, usually unglazed, which fires to a smooth, marblelike finish, and terra cotta, a yellow, red, or brown earthenware with no glaze, used for art sculpture, and similar purposes.

The firing of glazed ware is usually done in two steps; first the unglazed body is fired to give it strength, and then after the glaze is applied, it is refired at a lower temperature (except porcelains, where the second firing is at a higher temperature).

Absorption is determined by the presence of open pores or voids in the fired material into which water can penetrate; in general, the higher the firing temperature, the lower the absorption. Body color is determined mainly by raw-material purity. Strength depends on the porosity and also on the amount and type of glass and crystals developed in the body on firing. Translucency is obtained in products in which there is low porosity and little difference in index of refraction between the glass and crystals in the body. See **CLAY**.

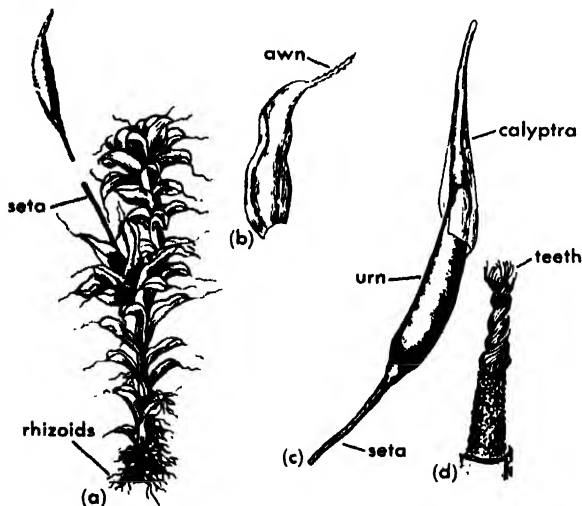


## Pottiales

An order of mosses. Bryologists are not in accord with regard to some of the families which have been classified in this order. The genera of mosses commonly considered in this order are numerous. They differ so strongly in some characteristics that it is difficult to include all of them in one description.

The plants vary in height from minute (up to 5 mm) to large (up to 10 cm), growing mostly in sods or tufts but occasionally scattered. The stems are erect, simple, or with dichotomous or fasciculate branching. The leaves are crowded, in several rows, and rarely 3-ranked. They vary in shape, but the majority are lanceolate to broadly ovate or obovate. The costa is strong, mostly percurrent or excurrent. The upper cells of the leaves are usually small, with thick, slightly to densely papillose walls.

The calyptra in Pottiales is cucullate. The erect seta varies from very short to elongate. Though the capsule is usually peristomate, it is cleistocarpous or gymnostomous in some species. The operculum



*Tortula ruralis*. (a) Entire plant. (b) Leaf. (c) Urn with calyptra. (d) Peristome. (From W. H. Welch, *Mosses of Indiana*, Ind. Dept. Conserv., 1957)

and peristome are present in some species of this order and absent in others. The peristome, when present, is composed usually of 16 teeth, although 32 are present in some species. The teeth are papillose, straight or spirally twisted, entire or cleft into 2-3 filiform divisions. In some species a basal membrane occurs at the base of the peristome. See MUSCI. [W.H.W.]

## Poultry production

The production of poultry embraces all phases in the life cycle of chickens, turkeys, ducks, and geese, but does not include game and other wild birds. The life cycle includes the incubation of the eggs, the growing of the stock for either meat- or egg-production, the egg-laying phase, and the breeding stage, when fertile eggs are produced for incuba-

tion. The subject will be discussed with particular reference to chickens, followed by a general discussion of turkeys, ducks, and geese.

**Incubation.** The practice of hatching eggs by artificial means rather than by the use of the hen dates back to the days of ancient Egypt. As a science, it has been developed primarily since the start of the twentieth century. The modern incubator is a scientific piece of equipment designed to give the developing embryo an ideal environment with automatic control of temperature and humidity, and to include labor-saving and safety-control devices.

The temperature required for incubation is 99.75°F. Variations of a degree or more from this optimum level will delay or advance the number of hours required for the chick to hatch. Uniformity of temperature is important to maintain a hatching schedule with eggs to go into the incubator at a specified time and the resulting chicks to be removed and shipped also at a prescribed time. Forced circulation of air around the eggs by the use of fans ensures an even distribution of the heat whether it be from electric coils, hot water pipes, or other sources. Thermostatic controls maintain the desired temperature.

Next in importance to heat control comes humidity regulation, the purpose of which is to conserve the moisture within the egg, yet permit some evaporation. Otherwise the embryo may drown. See EGG (FOWL). The requirements are not so exacting as those for temperature, a 60% relative humidity is satisfactory until the time of actual hatching, when it should be raised to 70%.

Providing ventilation to an incubator, thereby ensuring sufficient oxygen, is still a matter of judgment rather than scientific control. Most incubators have a manual control which is set to give sufficient ventilation when the machines are located in a warm, well-ventilated room. The room itself must receive serious consideration and be designed to maintain uniform conditions regardless of outside weather changes. Ventilation within the incubator is of greatest importance when the chicks are hatching and for a few hours following the hatch when they are drying off. Suffocation may easily occur with inadequate ventilation at this time. Because of the special requirements when the chicks are actually hatching, most incubators are designed to operate in two sections, one to hold the eggs for the first 18 days and the other as strictly a hatching compartment.

It is necessary to turn the eggs every 4 or 6 hours until the embryos are about ready to hatch. This prevents the embryo from adhering to the shell membranes. Automatic devices are timed to take care of this chore. Except for inspection to see that the equipment is functioning properly, an incubator does not require any measure of human control from the time the eggs are set until the eighteenth day of incubation.

The normal hatching expectancy is 80-90 chicks from each 100 eggs set. The number varies with efficiency of operation of the incubator, the ferti-

ity of the eggs, the season of the year, the diet and age of the breeding stock, and the length of time and environmental conditions under which the hatching eggs were held before being placed in the incubator.

Fertility of eggs generally will exceed 90%, but may be adversely affected if the male birds are sick, fight too much, become chilled, suffer from frozen combs or wattles, or are genetically inferior in breeding quality.

The season of the year affects hatchability of eggs more or less indirectly. Hatches may be poor in the fall when chickens normally molt, but a flock of young pullets that is not molting will produce hatching eggs of high quality at this season. Excessively hot or cold weather may adversely affect the hatching.

Age of breeding stock is a factor in the results obtained. Both the fertility of the eggs and the actual hatching of the fertile eggs decreases with advancing age of the stock.

The conditions under which eggs are held prior to incubation are important. In principle, eggs should be placed under incubation within 3 days after they are laid, but the life of the embryo may be maintained for a period of 10-20 days if the eggs are stored properly. The holding temperature must not exceed 65°F, and the humidity of the holding room should exceed 70%. Any undue loss of moisture from within the egg will be detrimental. The eggs will also need to be turned twice daily if held more than 1 week, in order to keep the yolk from adhering to the shell membranes.

The diet of the breeding hens must be complete in the nutrient factors required for the growth of the embryo. The vitamin requirements need particular attention. Hens will produce eggs on a diet quite inadequate for growth and such eggs will give low results when incubated. Finally, the breeding stock must be free of any diseases transmitted through the egg, such as pullorum disease (see *SALMONELLA*). Fortunately, the adult stock affected with this disease may be subjected to a blood test and carrier birds removed. Other diseases that lower the vitality of the stock affect the hatching of the eggs as will various parasites, both external and internal (see *CESTODA*; *MALLOPHAGA*; *NEMATODA*; *TREMATODA*).

**Baby chick industry.** Following the hatching of the eggs, there is a lapse of 3-5 days when the young chicks do not necessarily require feed or water. The yolk of the egg is drawn directly into the body cavity of the embryo at about the time the shell is broken and the chick prepares to emerge. This provides food for the new chick for a few days after hatching, a wise provision of nature to take care of the chick under adverse conditions. Man takes advantage of this protection period and uses this time to transport the chicks to all parts of the world. The commercial shipping of chicks started in the early part of this century and has advanced until millions of chicks are now hatched and shipped weekly. Many farmers no longer have any connection with the incubation of chicks; they

simply place their orders and receive the chicks when they want them. This development has made it possible for breeding operations to be more centralized with consequent improvement in the grade of chicks produced. In addition it has resulted in the growth of large farms requiring thousands of chicks at one time, an impossibility when the farmer had to depend on his own breeding flock to secure his supply of hatching eggs. Without the hatchery to produce chicks as needed, the commercial poultry industry as it exists today would be unknown.

**Brooding.** The brooding of chicks under artificial conditions dates back to the late 1800s when coal- and oil-burning stoves were used to furnish heat for brooding purposes. Many of these stoves have not changed much in design for more than 50 years which suggests that little advance has been made in the science of brooding. Electric and gas-burning stoves have been developed as well as large central heating systems which use either hot water or hot air, but the brooding of chicks in large units has not been too successful because of disease hazards. Batteries of wire cages have at times been used, but these also have not been too successful except for experimental groups where small numbers are kept. Although some farms may brood several hundred thousand chicks at one time, these chicks usually are started in small groups of 300-500 each.

The temperature requirement for baby chicks is 90°F for the first week, after which the temperature may be reduced 5 degrees weekly until the chicks are 6 weeks old. At that age they should be well feathered and able to withstand moderate temperatures above freezing providing they are protected from wind and rain.

In the brooding of chicks, human judgment is necessary in daily care of the stock. In climates or seasons where there are wide variations in day and night temperatures, the caretaker must assume more responsibility for the care and management of the chicks. No mechanical device can meet the situation. Overcrowding, overheating, or chilling, even for a few hours, can have very serious results.

Because chicks pick at anything in sight for the first few days after they hatch, the litter used to cover the floor must be something that will not be injurious. Sand has proved successful, but other materials, such as cut straw, wood shavings, peat moss, oat hulls, peanut shells, or most any nonpoisonous but moisture-absorbent product, may be used provided it is covered for the first few days with paper to prevent the chicks from eating it. The paper in turn may serve as a place to feed the chicks for the first 5 days or until they learn where the regular feed hoppers are located. Water also is a problem for the first few days because the chicks must learn where the water is located. In general the practice is to have 2 or 3 small containers for each group of chicks until they learn the location of the permanent water supply.

**Diet.** For the first few weeks of their lives, baby chicks should be given a diet that is complete in

Nutrient requirements for chickens in percentage or amount per pound of feed

	Starting chickens 0-8 weeks	Growing chickens 8-18 weeks	Laying hens	Breeding hens
Total protein, %	20	16	15	15
<b>Vitamins</b>				
Vitamin A activity, USP units	1200	1200	2000	2000
Vitamin D, ICU	90	90	225	225
Thiamine, mg	0.8	?	?	?
Riboflavin, mg	1.3	0.8	1.0	1.7
Pantothenic acid, mg	4.2	4.2	2.1	4.2
Niacin, mg	12	?	?	?
Pyridoxine, mg	1.3	?	1.3	1.3
Biotin, mg	0.04	?	?	?
Choline, mg	600	?	?	?
Folacin, mg	0.25	?	0.11	0.16
<b>Minerals</b>				
Calcium, %	1.0	1.0	2.25	2.25
Phosphorus, %	0.6	0.6	0.6	0.6
Salt, %	0.5	0.5	0.5	0.5
Potassium, %	0.2	0.16	?	?
Manganese, mg	25	?	?	15
Iodine, mg	0.5	0.2	0.2	0.5
Magnesium, mg	220	?	?	?

SOURCE: Natl. Acad. Sci.-Natl. Research Council. *Nutrient Requirements for Poultry*, Publ. 301, January, 1954. USP, U.S. Pharmacopeia; ICU, International clinical units.

all the needed proteins, minerals, and vitamins. The requirements in this respect have been fairly well established, more research having been done on the diet of baby chicks than on any other phase of the poultry business. The recommended allowances for the various components of the diet are shown in the table. Feed ingredients which supply the needed nutrients when mixed together in a ground form are known as mash and this is fed to the chicks as soon as possible after they are hatched. The amount of feed is not limited in any manner. Along with it, the chicks are fed a hard insoluble grit, such as granite, on the supposition that it aids in the utilization of the mash. Proof of this is not definite, however. Sometimes the mash is compressed into pellet form and fed in that manner, or the pellets may be reground to a crumbly consistency. The object in producing the pellet or its crumbly form is to improve the physical condition of the feed and thus to encourage a greater intake. Some feeding materials that in themselves are not palatable to chicks may be used to advantage if prepared in pellet form.

Before sufficient knowledge had been gained to mix a complete feed for chicks, the practice was to have chicks get out in the sunshine with a supply of grass available. Milk was used instead of water, and grains were fed in place of mash. Today research specialists in both the agricultural experiment stations and the feed industry ensure delivery of just what the chicks require for rapid growth and good health. The farmer's obligation is to see that the feed is placed before the chicks. Even that is purely mechanical because the feed is delivered in bulk and is stored in bins from which it moves directly into automatic feeders that run throughout the poultry house. Science has

made it possible for one man to take care of 20,000 or more chicks because the operations are mostly mechanical. Before all this could come about, the nutritive requirements had to be known so that the feeding could be as accurate and simplified as is now possible.

For young chickens to be sold for broiling or frying purposes, the diet remains the same as used for baby chicks. However, if the purpose in rearing is to produce pullets that are to be used later for egg production, the diet is changed at about the age of 8 weeks. Growth in chickens tends to divide itself into two periods. The first consists of a constantly increasing rate of gain in weight, week by week, the second by a constantly decreasing rate of gain in weight week by week. In chickens the break between the two periods comes at the age of about 8 weeks, although diet, temperature, breeding, and disease will tend to alter this to some extent. Therefore, starting at the age of 8 weeks, growing pullets are fed diets that are lower in protein as well as some of the other nutrients, as indicated in the table. Under practical field conditions, the changes are more radical than indicated in the table because vitamin D is not needed in the feed when the chickens have access to direct sunshine. Also, some of the other vitamin requirements may be met by having the young pullets on a good green range, and the diet may be altered accordingly. When the stock is being reared indoors or at a season when range facilities are unsatisfactory, the feed must be complete or the pullets will become unhealthy and growth will be stunted.

**Egg production.** The period of egg production is marked by a normal increase in the rate of lay from the time the first eggs are produced until the flock reaches a peak of 70-80% production, that is, 70-80 eggs from 100 pullets daily. This level generally is reached by the time the pullets are 8 months old. The production level will then gradually decline, and the peak will never again be reached in the life of the birds. Usually a complete feather molt occurs after egg production has continued for a year and, during this molting period, the rate of lay may be as low as 10%. After new feathers have been grown, egg laying is resumed for another year. This process is repeated year after year during the lifetime of the fowl, with a gradual decline in rate of egg production each succeeding year.

Several factors have effect on the egg-laying behavior of a flock of chickens. These include diet, environmental conditions, health and genetic constitution (see GENETICS).

After a suitable mixture of feeding materials has been established to meet the requirements, the resulting feed is fed in either mash, pellet, or crumbly form as indicated for chicks, usually by the use of automatic feeders. On farms where grain is grown, a portion of the diet may be supplied in whole grains, and thus the cost of the feed is reduced. Because grains are primarily a source of energy and are deficient in protein, minerals, and vitamins,

mins, a mash to supplement grain must be more concentrated in these factors than otherwise would be necessary. Commercial egg production is becoming more and more a speciality and the use of home-grown grains is fast losing its popularity. As was indicated in the case of incubation and the growing of chicks, the farmer no longer needs to have the technical knowledge to feed a flock of laying birds properly. He must see that the machinery functions, leaving the more technical problems to the hatcheryman and the feed dealer.

**Health.** Diseases of poultry have plagued the industry at all times and still present a problem, but the emphasis is now directed toward prevention of disease and maintenance of health, rather than obtaining a cure for a specific disease. An adult hen that is to perform economically as a producer of eggs must be a healthy individual. She must be reared free of parasites and not be retarded in growth by undue exposure to disease. Mortality in young stock has been brought fairly well under control by the virtual elimination of pullorum disease through blood testing of the parent stock, use of drugs to keep coccidiosis under control, vaccination for Newcastle disease, laryngotracheitis, and fowl pox, and feeding antibiotics to give greater resistance to other infections (*see* ANTIBIOTIC; COCCIDIA; NEWCASTLE DISEASE). Parasites may be kept under control by the judicious use of drugs and chemicals at the proper time, coupled with management practices that aid in keeping the incidence of infestation at a low level. Breeding for health and low mortality has taken on new significance and strains of chickens are available that have demonstrated resistance to disease. There are some diseases still to be investigated, chiefly those dealing with what is termed the "leukemia complex" or leucosis, resulting in paralysis and total disability.

**Environmental problems.** The effect of the environment on the behavior of chickens is in need of much research, because it includes the problems of controlling the hours and intensity of light, the temperature, humidity, ventilation, and floor space as they affect health and egg production. Density of the population is another factor. Light has been the subject of considerable study and egg production is known to be retarded if chickens are subjected to less than 12 hours of light daily. However, growing-stock does not show a similar response in relationship to growth rate. Optimum temperature for egg production has not been established, but chickens seem quite adjustable in this respect and able to perform satisfactorily in a wide range of temperature after they become acclimated. Humidity and ventilation requirements are still highly controversial. In this general area of knowledge there is urgent need for information regarding the physiology of the fowl which must be understood before the real effect of the environment can be properly determined.

**Heredity.** Although research has done much to advance knowledge of the nutritional needs of the chicken and has permitted the control of many dis-

cases, the genetic constitution of a fowl is highly important as well. Fortunately, geneticists have been working in this field, and there are now chickens that can grow rapidly and produce eggs at high rates when they are properly fed and managed. The crossing of breeds has resulted in marked improvement in the development of broilers and meat chickens. The dominance of certain desirable traits, such as the inheritance of silver plumage, has enabled the broiler grower to use males with this characteristic to produce offspring with light-colored feathers desired by the market. Sex of the chicks may be determined by the color of the down when the chicks hatch or by other markings, and the sexing of chicks by physical examination has replaced the practice of using different breeds to take advantage of known inherited sex differences.

Inheritance of egg-production qualities has proved to be a complicated problem. Starting originally with the use of males pedigreed from high-production hens, many flocks over the past 40 years have been brought to levels of production ranging from 200-225 eggs per bird per year. Considering that production in the farm flocks originally was about 150 eggs per bird, the advance has been well worth while. However, a plateau of egg production developed within strains after the use of pedigreed males over several generations, and it was not until strains were crossed that further improvement took place. Intensive inbreeding of different strains within the breed, followed by crossing, has brought about an increase in production up to levels exceeding 250 eggs per bird. This trend has had a tendency to place the breeding operations in the hands of a relatively few individuals or companies, because so many birds must be kept and tested for their ability to show improvement when crossed, that it is an expensive operation. The actual farm operator of the future will depend entirely on an outside organization to breed the type of chickens desired, and he will be responsible for only the physical care of the stock.

**Marketing.** The advent of the chain store system in merchandising brought with it a demand for large quantities of poultry products of uniform grade and quality. These were difficult to obtain when each farmer followed his own inclinations. As a result, poultry products are being produced and marketed under what is termed "integration," a system whereby the producer agrees to market his product through a prescribed channel with chicks from a definite source and fed a diet designed to give the best in quality. In some instances cooperative associations are serving the same purpose and give the farmer an element of control. The individual producer desiring to market products through private channels rather than an integrated system must keep the quality of his products at a peak level; otherwise the outlets will be greatly restricted. In particular, egg quality needs to be stressed because freshness is lost rapidly unless eggs are kept under refrigeration from the time they are laid until they reach the consumer.

**Economics.** Science has made the production of poultry and eggs a relatively simple matter in so far as the farmer or actual producer is concerned. Large-scale operations with thousands of birds now exist and, under improving management, the cost of production per unit of output is being brought lower and lower. This advantage in turn is passed on to the consumer who purchases poultry products in competition with other foods of equal nutritive value. With the price advantage in favor of poultry products, the outlook for the future seems good.

**Turkeys, ducks, and geese.** The general principles underlying the care of chickens apply equally well to turkeys, ducks, and geese even though the scientific requirements for these other species have not been so thoroughly determined. However, research in nutrition and disease control is progressing rapidly with some attention being given to genetics. Because turkeys, ducks, and geese are reared for the production of meat, they come into direct price competition with broilers and other meat products of the chicken. In due course and with sufficient research, there is every reason to believe that many of the advantages now enjoyed by the producer of chickens will be equally available to the farmer interested in turkeys, ducks, or geese, and that these species will be as commercialized as the chicken. See CHICKEN; DUCK; GOOSE; TURKEY; *see also* EGG PROCESSING; EMBRYOLOGY; OVUM; REPRODUCTION, ANIMAL.

[C.S.PL.]

**Bibliography:** See AGRICULTURAL SCIENCE (ANIMAL).

## Pound

A unit of mass in the English absolute system of units; also, a unit of force in the English gravitational system. The British standard of mass is the British Imperial Pound, of which a standard is preserved by the government. The United States standard mass is the avoirdupois pound, defined as the gravitational attraction, or weight, of 1/2.204622 kilogram.

A one-pound force is the weight of the British Imperial Pound at a standard location, that is, at any point where the acceleration of gravity is 32.174 ft/sec<sup>2</sup>. In terms of Newton's second law, there is a derived unit of mass, the slug, which is that mass to which a force equal to the standard pound will impart 1 ft/sec<sup>2</sup> acceleration. A one-pound force is equal to 4.4482 newton, and produces an acceleration of 32.174 ft/sec<sup>2</sup> when acting on a one-pound mass. See FORCE; MASS; MEASURE; UNITS, SYSTEMS OF; WEIGHT.

[L.N.]

## Poundal

A unit of force in the British absolute system of units. One poundal is the force which will impart to the British Imperial Pound mass an acceleration of 1 ft/sec<sup>2</sup>. The foot is 1/3 of the Imperial Standard Yard. One poundal is 0.13825 newton. See FORCE.

[G.E.P.]

## Pour point

The lowest temperature at which an oil will pour when cooled under prescribed test conditions. Because petroleum oils are complex mixtures which become plastic solids when cooled, several solidification temperatures are used, each defined by a definite test procedure. The cloud point is the temperature at which a solid phase separates from solution. For a grease, the dropping point is the temperature at which the plastic solid becomes sufficiently fluid to flow through an orifice. For waxes, the melting point is the temperature at which the material becomes fluid enough to drop from the test thermometer; the congealing point is the temperature at which a sample wetting the thermometer appears to congeal and hence rotate with the thermometer. See OIL ANALYSIS. [M.SO.]

## Powder metallurgy

A process of the metallurgy industry involved with the production of finely comminuted metal powders, and of metal objects directly from these powders.

The products are usually finished parts such as gears or cams. The technique employed consists essentially of subjecting the metal powders to pressure and heat. The heat treatment, called sintering, is performed at some temperature below the fusion point of the main constituents of the products. Instead of pure metal powders, alloy powders may be used singly or as mixtures. Also, metal powders may be used in mixtures with metallic compounds or nonmetallic components. Powder metallurgy thus permits the production of metallic, or metal-like, bodies of many shapes without the use of orthodox metallurgy practices such as melting and casting.

**Industrial applications.** Since many refractory metals have such high melting points that conventional melting and casting is difficult, powder metallurgy is the ideal method of producing such products as tungsten for filaments. Metal combinations in which the characteristics of each constituent are retained are of particular interest for certain electrical applications, and they can be produced by powder metallurgy methods. For instance, heavy-duty electrical contacts and welding electrodes combine a skeleton of refractory metal, highly resistant to abrasion and arcing, with a second metal of low melting point and high conductivity. Alloying between the constituents is negligible so that the original properties of the individual metals are preserved.

Other examples are the manufacture of cemented carbide high-speed cutting tools, cermets, and dispersion alloys. Cermets consist of a predominant nonmetallic, or ceramic, constituent and a metallic binder phase. Dispersion alloys contain minute nonmetallic particles dispersed in a metallic matrix. They find important applications in nuclear reactor components such as fuel elements, which consist of combinations of uranium oxide and binder metals,



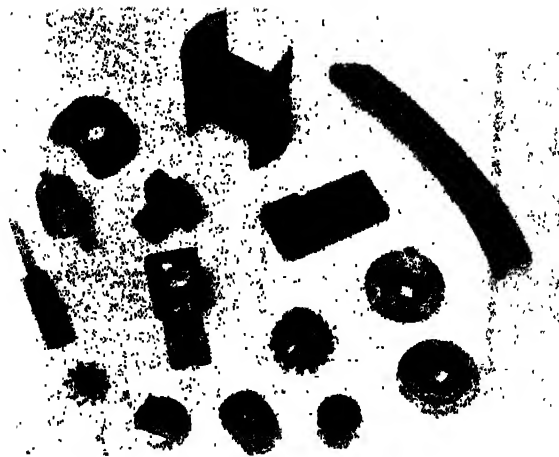


Fig. 1. Assortment of structural powder metallurgy parts of complex shape. (Metal Powder Industries Federation)

and control or moderator elements, which contain fine dispersions of boron and other neutron-capturing elements in aluminum, stainless steel, or zirconium matrix metals. See CERMET.

Of great interest is the production of small metal parts such as gears, cams, and other components for machines and instruments requiring closely controlled dimensions and properties. These parts can often be produced at lower cost by powder metallurgy than by other methods of shaping metals. The parts may be steel, brass, or alloys of iron with copper, nickel, or chromium.

The development of metallic bodies of closely controlled porosity has made possible so-called self-lubricating bronze and iron-base bearings, which are initially impregnated with oil and used in places which are inaccessible to external lubrication. Porous metal is also used effectively in oil-pump gears, metal filters, and diaphragms.



Fig. 2. Press tools and sections of brass-infiltrated iron part shown in foreground. Tools include upper punch (left), die (center), lower punch (right) and core rod (lower right). The thin brass infiltrant blanks (lower center) are compacted with the same tools. (Metal Powder Industries Federation)

Current collector brushes in electrical machinery are laminated metal powder products. Copper-lead bearings whose constituents are not miscible in the liquid or solid state are typical of metal powder alloys of unusual compositions. Other applications of metal powders include dental alloys, chemical reagents, catalysts, explosives and pyrotechnics, solders, brazing agents, coatings, cement additives, pigments, and flame cutting agents.

#### **Powder sources, production, characteristics.**

Virtually all metals and metalloids, and many of their alloys and compounds, are available in powder form. They may be obtained from ores, salts, and other compounds or from bulk metals and alloys. The methods of production may be classified into mechanical, physical, chemical, and electrical processes. They include crushing, milling, machining, graining, atomizing, condensation, reduction, precipitation, displacement, electrodeposition, diffusion alloying, and alloy disintegration. The characteristics of the powders depend to a great extent upon the specific manufacturing process. The size of the individual particles can be made to vary over a wide range, from granules of  $\frac{1}{16}$  in. and coarser, down to near colloidal size, perhaps less than 1 micron in diameter. Concomitant with a decrease in size, the cumulative surface area of the particles increases rapidly to very large dimensions. This has an extremely important bearing on the properties of the powder, its behavior during processing into solid bodies, and the ultimate properties of these products. Another factor affecting the processing behavior and final properties is the shape of the individual particles. This, too, depends on the method of producing the powder. Particles of mechanically comminuted powders are generally solid in structure, angular in form, and may be equiaxed, elongated, or flaked in shape. Particles of physically produced powders are also generally solid in structure but mostly spheroidal in shape. Particles of chemically produced powders are always porous, often very light and fluffy, and usually consist of a multitude of individual crystallites. Electrolytic powder particles vary in shape and size depending on the process used. Fused salt electrolysis produces particles resembling those obtained by the chemical processes because they require chemical after-treatment for removal of salt residues and for reduction. Aqueous solution electrolysis produces either solid angular monocrystalline particles resulting from the mechanical comminution of brittle cathode deposits, or leafy dendritic aggregates of crystallites from direct electrolytic deposits.

Some metal powders can be produced by only one of these methods, limiting their properties to those peculiar to that method. Most industrially important metals can be produced in powder form by more than one method, thus permitting selection of the powder with the most suitable characteristics for the specific end product. Brass powders are produced exclusively by atomization, but copper



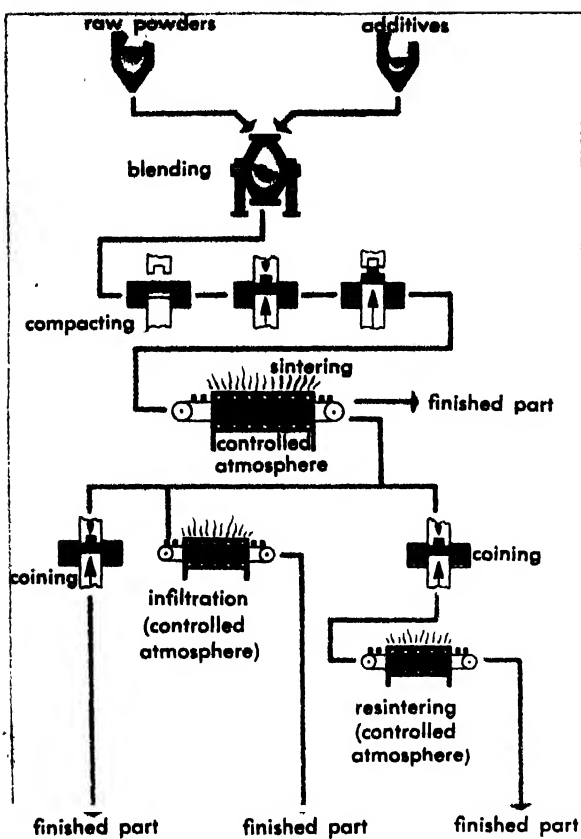


Fig. 3. Flow chart for fabrication of parts by powder metallurgy. (Metal Powder Industries Federation)

powders are produced by chemical precipitation, reduction, and electrolysis. Stainless steel powders can be produced effectively only by chemical disintegration of a specially prepared sheet metal, but iron powder can be produced by atomization, reaction with solid or gaseous reducing agents, electrolysis, or decomposition of iron carbonyl.

**Powder conditioning and processing technique.** The powdered raw materials are cleaned, dry stored, and tested. Usually they are mixed with lubricating agents to facilitate pressing. Very brittle metal, or compound, powders require careful admixing of organic binders to improve coherency during and after compaction. Powders for alloys or composite structures are prepared by blending desired proportions of the ingredients. This blending operation is carried out under various conditions and usually with dry powders. Often blending is combined with ball milling. Precautions must be taken to prevent contamination, deterioration, or ignition of the powders during storage, handling, or blending.

Most commonly, the powders are pressed to form coherent masses and the resulting compacts are sintered. The pressing operation is generally carried out at ambient, but sometimes elevated, temperatures. The powders are confined in closed dies, and heavy hydraulic or quick-acting mechanical tableting presses provide the compacting force.

Rolling of powders into strips is another method of compaction. In special cases, such as the manufacture of porous filters, compression of the powder prior to sintering is omitted entirely. Molds that allow for shrinkage are used either before or during sintering to impart the desired shape to the powder mass. When pressing is done at sufficiently elevated temperatures, the sintering operation takes place under pressure. The application of heat during or after the compression of the powder produces a structure with physical properties comparable to those of similar materials obtained by fusion. The sintered structure, however, possesses a fine porosity which has an adverse effect mainly on ductility. The porosity results from incomplete densification during compression and from gas evolution during heating. It cannot be completely eliminated by the sintering phenomena of diffusion, recrystallization, grain growth, and shrinkage.

Although the properties of the object after sintering are frequently adequate for the finished product, it is sometimes necessary to improve them by further operations, such as additional cold or hot compression, metal working, heat-treatments, or the impregnation of the pores with a lower-melting metal. Certain other finishing steps may also be required, including machining, plating, buffing, sanding, grinding, or barrel tumbling. A number of applications require joining operations. Brazing, soldering, or welding of these parts onto other metal bases is common practice in the hard metal and refractory metal industries. Finishing and joining operations must frequently be adapted to the specific properties of these articles. Thus, care must be taken in machining because of the porosity. Plating methods must be adjusted to prevent corrosion caused by the porosity, and special fluxes or inert or reducing gaseous media must be used to prevent excessive oxidation during brazing and welding at high temperatures.

**Critical evaluation.** A comparison of powder metallurgy with more orthodox metallurgical methods reveals the advantages and limitations of the method. This may well serve as a guide for the engineer and manufacturer confronted with the problem of selecting the most suitable production method for a particular job.

Powder metallurgy does not make up a preponderant segment of the metallurgical industries. In spite of the accelerated growth of its different branches, powder metallurgy has remained a limited and specialized field. The reasons for this are both economic and technical in nature. Powdered starting materials, with the exception of most iron powders, are more expensive than other raw materials. Tools and dies must be durable, and usually return their cost only when many thousands of the same part are produced. Special tools are required for the forming of complex shapes because the powder does not flow readily into lateral protrusions. Powders of high specific volume require long compression strokes, which in turn impose slow

production rates. Also, the large surfaces of the powders are prone to excessive gas absorption, leading to brittleness of the end product.

Yet powder metallurgy does have a potential for further expansion and development of new applications. Beryllium products are made exclusively by powder metallurgy. Additional, and perhaps spectacular advances, may be expected in the electronics, nuclear, and rocket fields, where powder metallurgy dispersion alloys, refractory metals, and metal-ceramics with unusual properties are being developed. See METALLURGY; SINTERING. [C.G.CO.]

*Bibliography:* C. G. Goetzl, *Treatise on Powder Metallurgy*, 3 vols., 1949-1952.

## Power

The time rate of doing work. Like work, power is a scalar quantity, that is, a quantity which has magnitude but no direction. Some units often used for the measurement of power are the watt (1 joule of work per second) and the horsepower (550 foot-pounds of work per second). See HORSEPOWER; WATT; WORK.

**Usefulness of the concept.** Power is a concept which can be used to describe the operation of any system or device in which a flow of energy occurs. In many problems of apparatus design, the power, rather than the total work to be done, determines the size of the component used. Any device can do a large amount of work by performing for a long time at a low rate of power, that is, by doing work slowly. However, if a large amount of work must be done rapidly, a high-power device is needed. High-power machines are usually larger, more complicated, and more expensive than equipment which need operate only at low power. A motor which must lift a certain weight will have to be larger and more powerful if it lifts the weight rapidly than if it raises it slowly. An electrical resistor must be large in size if it is to convert electrical energy into heat at a high rate without being damaged.

**Electrical power.** The power  $P$  developed in a direct-current electric circuit is  $P = VI$  where  $V$  is the applied potential difference and  $I$  is the current. The power is given in watts if  $V$  is in volts and  $I$  in amperes. In an alternating-current circuit,  $P = VI \cos \phi$ , where  $V$  and  $I$  are the effective values of the voltage and current and  $\phi$  is the phase angle between the current and the voltage. See ALTERNATING CURRENT.

**Power in mechanics.** Consider a force  $F$  which does work  $W$  on a particle. Let the motion be restricted to one dimension, with the displacement in this dimension given by  $x$ . Then by definition the power at time  $t$  will be given by

$$P = dW/dt$$

In this equation  $W$  can be considered as a function of either  $t$  or  $x$ . Treating  $W$  as a function of  $x$  gives

$$P = \frac{dW}{dt} = \frac{dW}{dx} \frac{dx}{dt}$$

Now  $dx/dt$  represents the velocity  $v$  of the particle, and  $dW/dx$  is equal to the force  $F$ , according to the definition of work. Thus

$$P = Fv$$

This often convenient expression for power can be generalized to three-dimensional motion. In this case, if  $\phi$  is the angle between the force  $F$  and the velocity  $v$ , which have magnitudes  $F$  and  $v$ , respectively,

$$P = \mathbf{F} \cdot \mathbf{v} = Fv \cos \phi$$

This equation expresses quantitatively the observation that if a machine is to be powerful, it must run fast, exert a large force, or do both. [P.W.S.]

## Power amplifier

The final stage in multistage amplifier circuits, such as audio amplifiers and radio transmitters, designed to deliver appreciable power to the load.

Power amplifiers may be called upon to supply power ranging from a few watts in an audio amplifier to many thousands of watts in a radio transmitter. In audio amplifiers the load is usually the dynamic impedance presented to the amplifier by a loudspeaker, and the problem is to maximize the power delivered to the load over a wide range of frequencies. The power amplifier in a radio transmitter operates over a relatively narrow range of frequencies with the load essentially a constant impedance.

The mode of operation of power amplifiers is denoted by Class A, AB, B, and C. Class C operation is limited to radio frequencies with a tuned load. The other classes may be used for audio and high-frequency operation. For discussion of the modes of operation, see AMPLIFIER.

**Class A power amplifiers.** Class A operation is used when the amount of power transferred to the load is relatively small, say, less than 10 watts. The amount of harmonic distortion introduced into the load voltage can be kept small by using tubes with nearly linear characteristics and restricting the range of operation to a small displacement from the operating point. This class of operation has relatively little use because the plate-circuit efficiency (the efficiency of a power amplifier) is low. The maximum possible efficiency is 50%. However, for the usual operating conditions and standard vacuum tubes, the efficiency is on the order of 10%. If the power amplifier were required to deliver 10 watts with 10% efficiency the tube would have to be capable of dissipating an average power of 100 watts. Furthermore, the power supply must be capable of supplying the power dissipated as heat by the tube plus the useful power delivered to the load. This poses an unnecessary burden upon the power supply. Other classes of operation have a

higher plate-circuit efficiency and are therefore used for higher-power applications.

**Class AB power amplifier.** An improvement in the plate-circuit efficiency can be had by using Class AB operation. However, while a Class A amplifier can be operated single-ended (one output tube), a Class AB amplifier must be operated push-pull. In Class AB operation the tube current does not flow for the complete cycle of the input voltage. In a single-ended circuit this would introduce excessive distortion. See PUSH-PULL AMPLIFIER.

**Class B operation.** This class is often used for the power amplifier in an audio amplifier. The amplifier in this class must be a push-pull circuit. Theoretically, with ideal tubes, the Class B amplifier can have a plate-circuit efficiency of 78.5%; practically, the efficiency is on the order of 50%, an appreciable improvement over that of Class A operation. Another advantage of Class B push-pull operation is that the average currents of the two tubes flow in opposite directions through the primary winding of the output transformer, resulting in no average magnetization of the core. This allows the use of smaller cores with savings in size and weight.

The load is transformer-coupled to the two tubes operating in push-pull. For maximum power transfer the dynamic load impedance presented to the amplifier should equal the conjugate of the output impedance of the amplifier when considered as an equivalent generator. In practice this is modified somewhat, because the harmonic distortion is dependent upon the dynamic resistance. The choice of this load is therefore determined by the amount of harmonic distortion that can be tolerated.

The use of more sophisticated circuitry than that considered in an elementary presentation of a push-pull amplifier operating in Class B can produce nearly distortionless power amplification. This is of prime importance in the final amplifier stages of a high-fidelity audio amplifier.

Power amplifiers can operate in Class B<sub>2</sub> with an appreciable amount of grid current flowing for a small portion of the cycle of an input sinusoidal signal. This imposes additional requirements upon the driving circuit of the amplifier. If the equivalent circuit of the driver has too large an equivalent output impedance, the flow of grid current through this impedance will cause a distortion of the grid waveform. This problem is usually encountered in the design of driver circuits for Class B amplifiers operating in the radio-frequency region. Audio operation is usually restricted to Class B<sub>1</sub> operation, because the usual form of phase-inverter circuitry has a large output impedance. In radio-frequency operation the transformer phase inverter can be used with tuned circuitry and air-core or powdered-iron slug coils, because the operation is essentially at one frequency.

**Class C operation.** Because the plate current flows for less than one-half cycle of the input sinusoidal signal, this class of operation is restricted to

radio-frequency operation where a tuned load is employed. The load is usually the input impedance of an antenna or of an antenna matching network. The load voltage will be nearly sinusoidal, even though the current in the tube flows in pulses, because of the relatively sharp tuning of the load. This phenomenon allows the amplification of large amounts of power at plate-circuit efficiencies as high as 80%. This is extremely important for applications requiring delivery of large amounts of power to the load.

The driving source must usually be called upon to deliver power to the grid circuit of the power amplifier, in many cases as much as 10% of the power delivered to the load. This requirement is not excessive. A Class B power amplifier can be used in the grid-driving circuit to obtain a quite efficient combination of driver and final amplifiers.

[H.F.K.]

*Bibliography:* J. D. Ryder, *Engineering Electronics*, 1957; S. Seely, *Electron-Tube Circuits*, 2d ed., 1958.

## Power factor

The ratio of watts average (or active) power to the apparent power of an alternating-current circuit (see ALTERNATING-CURRENT CIRCUIT THEORY). By definition, and of general application,

$$\text{Power factor (pf)} = \frac{\text{watts average power}}{\text{rms volts} \times \text{rms amperes}}$$

which is the ratio of instrument readings. A wattmeter indicates average power and electrodynamicometer or iron-vane instruments show rms voltage and current. For the steady-state ac circuit under sinusoidal voltage and current,  $\text{pf} = \cos \theta$ , where  $\theta$  is the phase angle between the voltage and current. This definition is restricted to sine waves of the same frequency. See WAVEFORM, NONSINUSOIDAL for the more general case.

[B.L.R.]

## Power plant

A means for converting stored energy into work. Stationary power plants such as electric generating stations are located near sources of stored energy, such as coal fields or river dams, or are located near the places where the work is to be performed, as in cities or industrial sites. Mobile power plants for transportation service are located in vehicles, as the gasoline engines in automobiles and diesel locomotives for railroads. Power plants range in capacity from a fraction of a horsepower to 500,000 kw in a single unit (Table 1). Large power plants are assembled on location from components made by different manufacturers. Smaller units are manufactured.

Most power plants convert part of the stored raw energy of fossil fuels into kinetic energy of a spinning shaft. Some power plants harness nuclear energy. For transportation, the plant may produce a propulsive jet instead of the rotary motion of a shaft. Other sources of energy, such as winds, tides,

Table 1. Representative design and performance data on power plants

Type	Unit size range, kw	Fuel*	Plant weight, lb/kw	Plant volume, ft <sup>3</sup> /kw	Heat rate, Btu/kw-hr
Central station					
hydro	10,000-100,000				
steam	10,000-500,000	CGN		20-50	8,500-15,000
diesel	1,000-5,000	DG			10,000-15,000
Industrial (by-product) steam	1,000-25,000	CGW		50-75	4,500-6,000
Diesel locomotive	1,000-5,000	D	100-200	2-3	10,000-15,000
Automobile	25-300	G'	5-10	0.1	15,000-20,000
Outboard	1-50	G'	2-5	0.1-0.5	15,000-20,000
Truck	50-300	D	10-20		12,000-18,000
Merchant ship, diesel	5,000-10,000	D	300-500		10,000-12,000
Naval vessel, steam	25,000-100,000	DN	25-50		12,000-18,000
Airplane, reciprocating engine	1,000-3,000	G'	1-3	0.05-0.10	12,000-15,000
Airplane, turbojet	3,000	D'	0.2-1		15,000-18,000

\* C, coal; D, diesel fuel; D', distillate; G, gas; G', gasoline; N, nuclear; W, waste

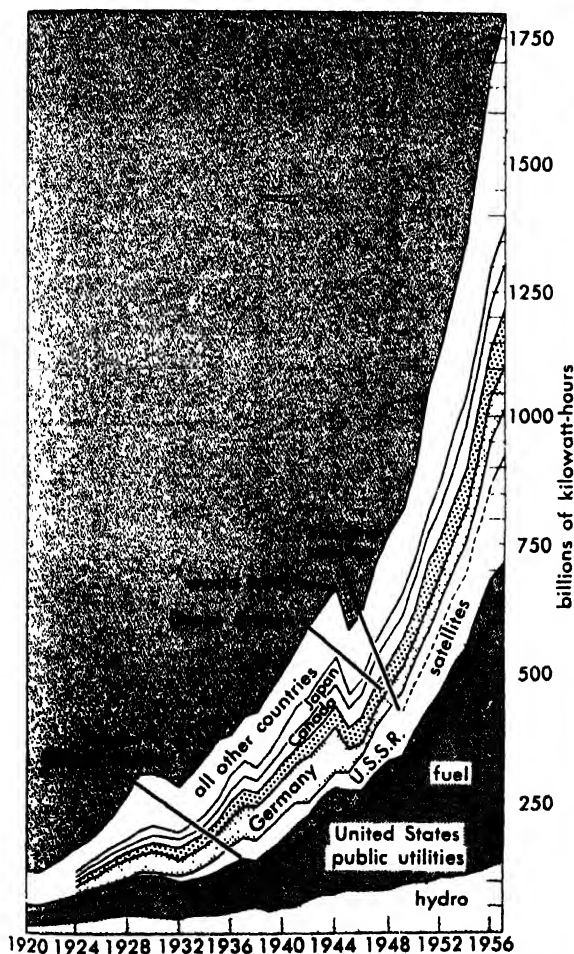


Fig. 1. Annual world production of electric energy. Generation by all agencies including industry for its own use. (Edison Electric Institute)

waves, and solar radiation, are of negligible commercial significance in the generation of power despite their tremendous magnitudes. Table 2 shows the scope of United States power plant capacity. About a third of the world's electric power is gen-

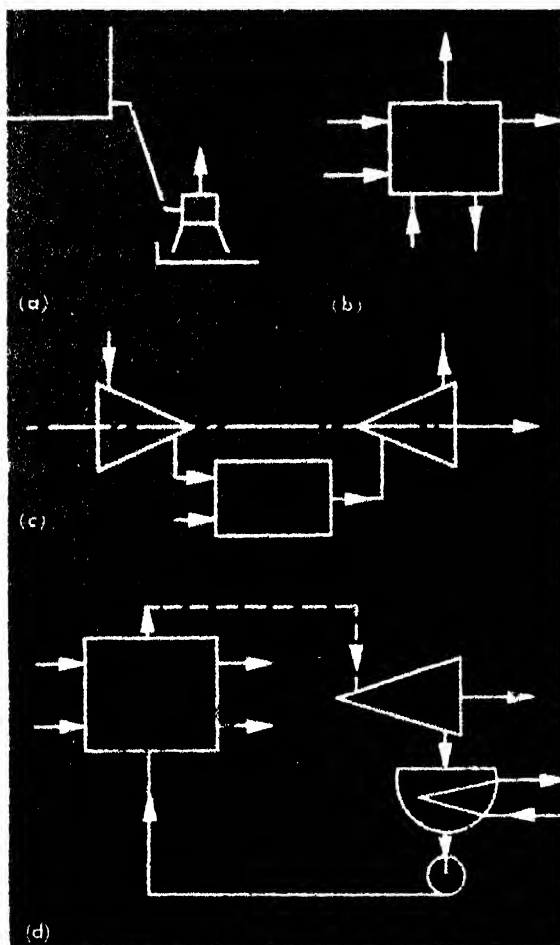


Fig. 2. Rudimentary flow diagrams for power plants: (a) hydro, (b) internal combustion, (c) gas-turbine, (d) condensing steam.

erated by United States public utility companies (Fig. 1). These data reflect the dominant position of fuel-generated power both for stationary service and for the propulsion of land-, water-, and air-borne vehicles.

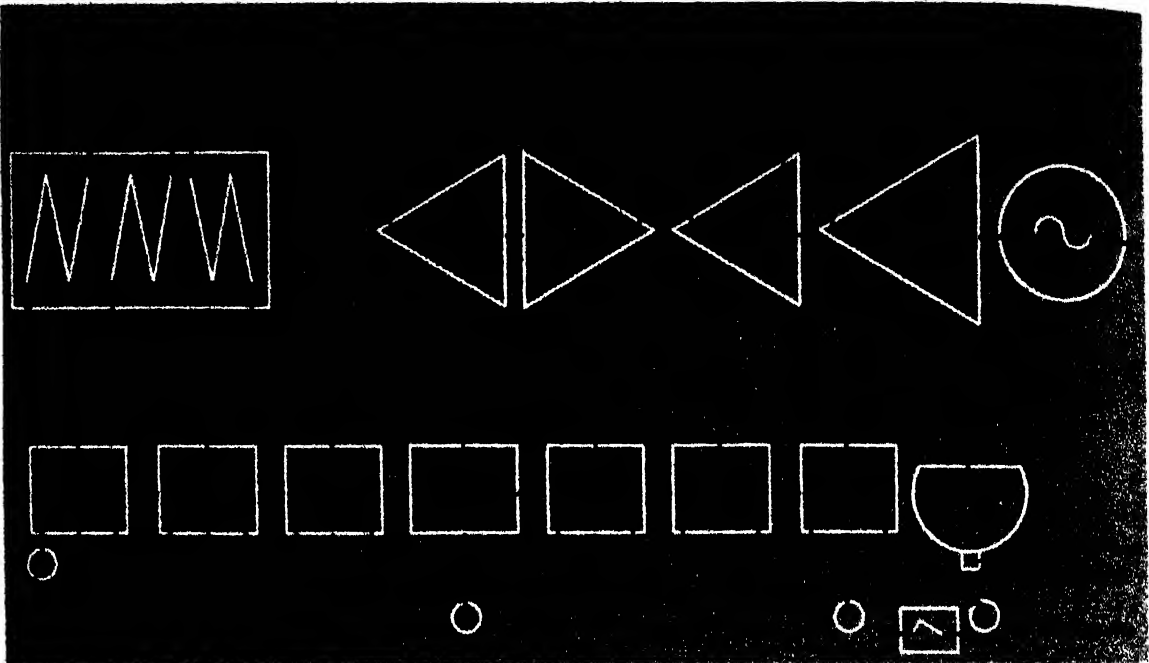


Fig. 3. Simplified heat balance for a supercritical, double reheat cycle. Net generation, 119,000 kw; plant heat rate, 8600 Btu kw-hr.

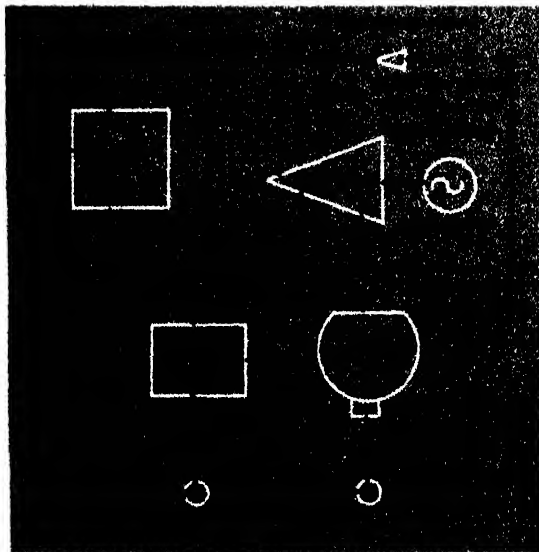


Fig. 4. Heat balance for a by-product industrial power plant delivering both electric energy and process steam.

Rudimentary flow- or heat-balance diagrams for important types of practical power plants are shown in Fig. 2 (see HEAT BALANCE). Many variations are incorporated for the control of efficiency, weight, space, flexibility, reversibility, reliability, life, investment, and operating costs. Figure 3 shows a heat-balance diagram for a modern steam-electric central station. Figure 4 is a similar diagram for a by-product-type industrial plant which

Table 2. Approximate installed capacity of United States power plants (1957)

Plant type	Capacity, millions of kw
Electric central stations	130
Industrial	30
Agricultural	50
Railroad	80
Marine, civilian	20
Aircraft, civilian	20
Military establishment	1000±
Automotive	6000±
Total	7000±

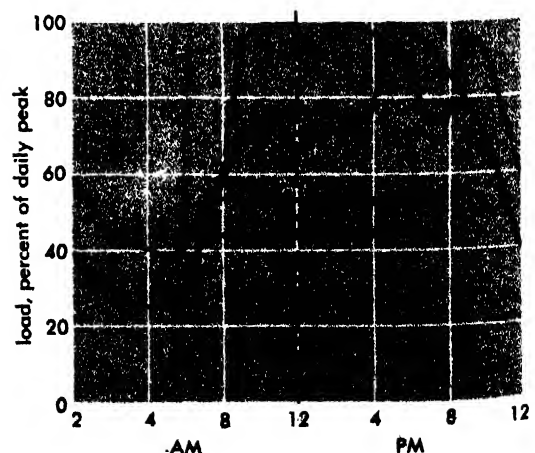


Fig. 5. Selected daily load curves for an urban utility plant.

**Table 3. Range of capacity factors for selected power plants**

Public-utility systems, in general	50-70
Chemical or metallurgical plant, three-shift operation	80-90
Seagoing ships, long voyages	70-80
Seagoing ships, short voyages	30-40
Airplanes, commercial	20-30
Private passenger cars	1-3
Main-line locomotives	30-40
Interurban buses and trucks	5-10

has the double purpose of generating electric power and simultaneously delivering heating steam by extraction or exhaust from the prime mover.

**Plant load.** There is no practical way of storing the mechanical or electrical output of a power plant in the magnitudes encountered in power-plant applications. The output must be generated at the instant of its use. This results in wide variations in the loads imposed upon a plant. The capacity, measured in kw or hp, must be available when the load is imposed. Much of the capacity may be idle during extended periods when there is no demand for output. Hence much of the potential output, measured as kw-hr or hp-hr, cannot be generated because there is no demand for output. This greatly complicates the design and confuses the economics of power plants. Kilowatts cannot be traded for kilowatt-hours, and vice versa.

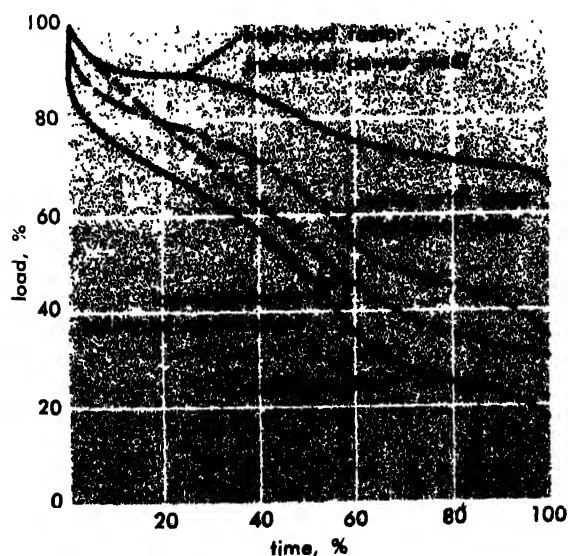
The ratio of average load to rated capacity or peak load is expressed as:

$$\text{Capacity factor} = \frac{\text{average load for the period}}{\text{rated or installed capacity}}$$

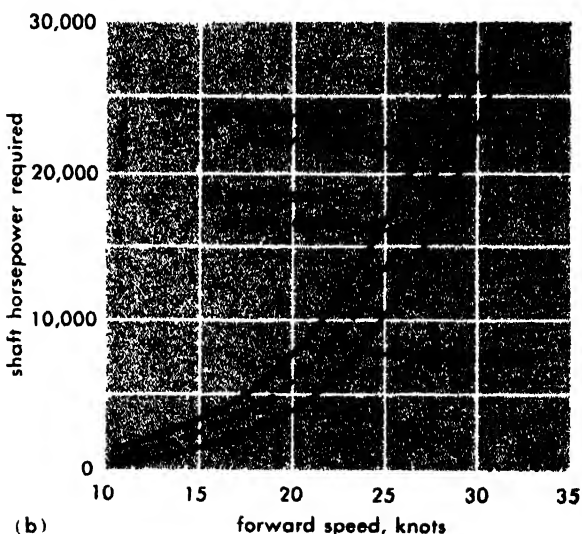
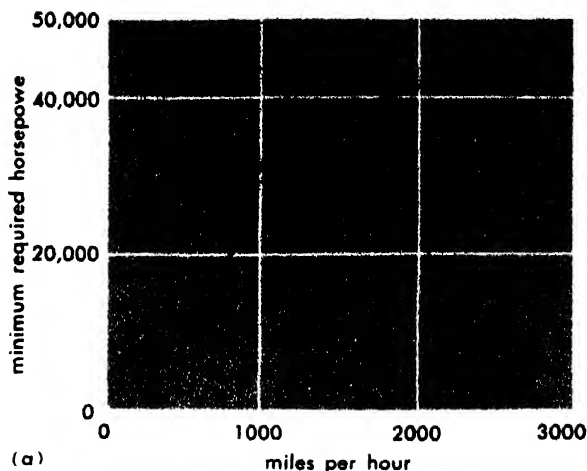
and

$$\text{Load factor} = \frac{\text{average load for the period}}{\text{peak load in the period}}$$

Some experienced capacity factors are given in Table 3.



**Fig. 6. Annual load-duration curves for selected stationary power plants.**



**Fig. 7. (a) Minimum power required to drive a 50-ton well-designed airplane in straight level flight at 30,000 ft altitude. (b) Power required to drive a ship, showing effect of fouling.**

Variations in loads can be conveniently shown on graphical bases as in Figs. 5 and 6 for public utilities and in Fig. 7 for air and marine propulsion. Rigorous definition of load factor is not possible with vehicles like tractors or automobiles because of variations in the character and condition of the running surface. In propulsion applications, power output (kw or hp), may be of secondary significance and performance may be based on tractive effort, drawbar pull, thrust, climb, and acceleration.

**Plant efficiency.** The efficiency of energy conversion is vital in most power plant installations. The theoretic power of a hydro plant in kw is  $QH/11.8$  where  $Q$  is the flow in cubic feet per second and  $H$  is the head at the site in feet. Losses in headworks, penstocks, turbines, draft tubes, tail-race, bearings, generators, and auxiliaries will reduce the salable output 15-20% below the theoretic in modern installations. The selection of a particular type water wheel depends on experience with wheels at the planned speed and on the lowest



Table 4. Cost of selected stationary electric power plants (Investor-owned and business-managed)

	Steam, central station		Hydro		Steam, industrial	
	Large	Small	Large	Small	Large	Small
Plant capacity, kw	500,000	50,000	200,000	20,000	10,000	1,000
Investment, \$ per kw	150	250	200	300	175	225
Fuel cost, cents per million Btu	20	30			30	40
Cost of power, mills per kw-hr						
Total cost	4.7	13.0	3.5	6.5	5.2	9.0
Carrying cost on investment	2.5	5.7	3.0	5.7	3.3	5.0
Production cost, total	2.2	7.3	0.5	0.8	1.9	4.0
Fuel	1.9	4.3			1.3	2.0
Labor, maintenance, supplies, supervision	0.3	3.0	0.5	0.8	0.6	2.0

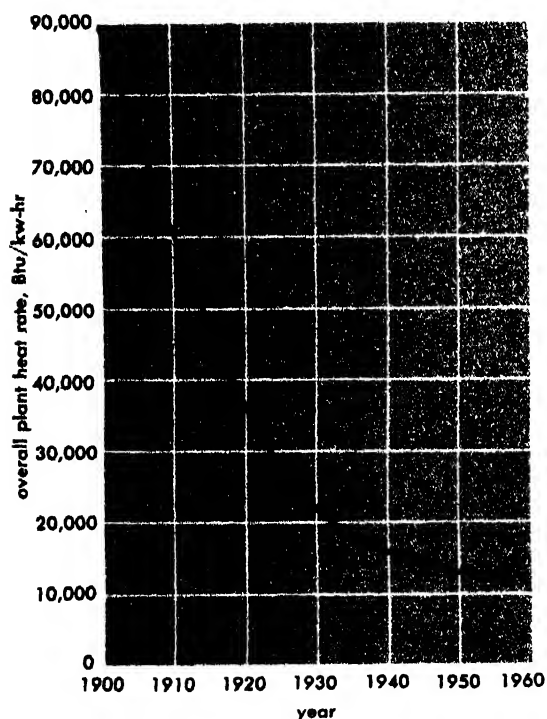


Fig. 8. Thermal performance of fuel-burning electric-utility power plants in the United States.

water pressure in the water path (see CAVITATION). Runners of the reaction type (high specific speed) are suited to low heads (below 500 ft) and the impulse type (low specific speed) to high head service (about 1000 ft). The lowest heads (below 100 ft) are best accommodated by reaction runners of the propeller or the adjustable blade types. Mixed-pressure runners are favored for the intermediate heads (50–500 ft). Draft tubes, which permit the unit to be placed safely above flood water and without sacrifice of site head, are essential parts of reaction unit installations.

With thermal power plants there are the basic limitations of thermodynamics which fix the effi-

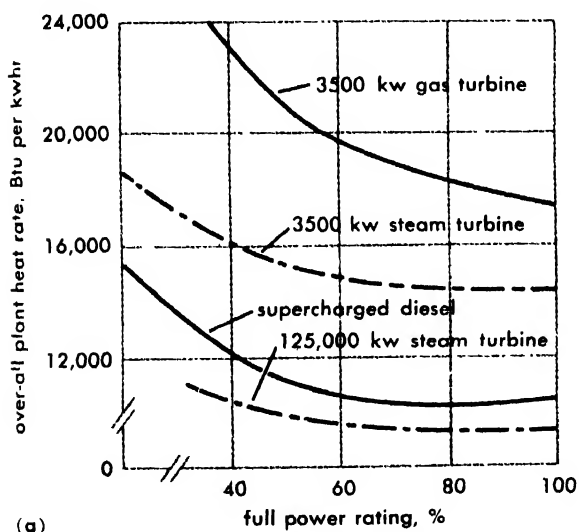
ciency of conversion for heat into work. The cyclic standards of Carnot, Rankine, Otto, Diesel, and Brayton are the usual criteria on which heat-power operations are variously judged. Performance of an assembled power plant, from fuel to net salable or usable output, may be expressed as thermal efficiency (%); fuel consumption (lb, pt, or gal per hp-hr or per kw-hr); or heat rate (Btu supplied in fuel per hp-hr or per kw-hr). American practice uses the high or gross calorific value of the fuel for measuring heat rate or thermal efficiency and differs in this respect from European practice, which prefers the low or net calorific value. Tables 1 and 3 give performances for several selected operations. Figure 8 reflects the improvement in fuel utilization of the United States electric power industry since 1900. Figure 9 shows variation in heat rate with load for several types of stationary plants and marine power plants.

In scrutinizing data on thermal performance, it should be recalled that the mechanical equivalent of heat (100% thermal efficiency) is 2545 Btu/hp-hr and 3413 Btu/kw-hr. Modern steam plants in large sizes (75,000–500,000 kw units) and internal combustion plants in modest sizes (1000–5000 kw) have little difficulty in delivering a kw-hr for less than 10,000 Btu in fuel (34% thermal efficiency). Lowest fuel consumptions per unit output (8500–9000 Btu/kw-hr) are obtained in condensing steam plants with the best vacua, regenerative-reheat cycles using eight stages of extraction feed heating, two stages of resuperheat, primary pressures of 5000 psi (supercritical) and temperatures of 1150°F. An industrial plant in which electric power is generated as a by-product of the process steam load can have a thermal efficiency of 5000 Btu/kw-hr.

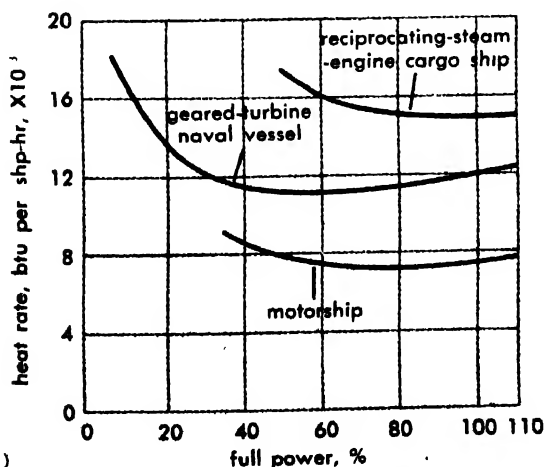
The atomic power plant substitutes the heat of fission for the heat of combustion, and the consequent plant differs only in the method of preparing the thermodynamic fluid. It is otherwise similar to the usual thermal power plant. Low reactor temperatures lead to the overwhelming preference for

steam-turbine rather than gas-turbine cycles. When fluid temperatures can be had above 1200°F, the gas-power cycle will receive more favorable consideration. Otherwise the atomic power plant is essentially a low pressure, low temperature steam operation (less than 1000 psi and 600°F).

**Power economy.** Costs are a significant, and often controlling, factor in any commercial power-plant application. Average costs have little significance because of the many variables, especially load factor. Some plants are short-lived and others long-lived. For example: in most automobiles, which have short-lived power plants, 100,000 miles and 3000 hrs constitute operating life; diesel locomotives, which will run 20,000 miles a month with complete overhauls every few years, and large seagoing ships, which register 1,000,000 miles of travel and still give excellent service after 20 years of operation, have long-lived plants; electric central stations of the hydro type remain in service 50 years, and steam plants run round the clock and upward of 8000 hours a year with complete reliability even



(a)



(b)

Fig. 9. Comparison of heat rates of selected types of (a) stationary power plants and (b) marine propulsion plants.

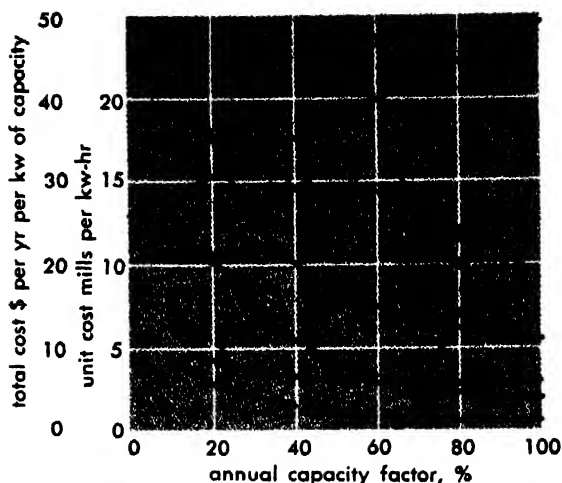


Fig. 10. Representative costs of power from a large steam-electric generating station.

when 25 years old. Such figures greatly influence costs. Furthermore, costs are open to wide differences of interpretation.

Costs are reflected in the curves of Fig. 10 for a modern steam-electric, investor-owned, central station in the eastern United States. Fixed costs are based on \$150 per kw investment with 12% annual carrying charges. If such a plant were government-financed, the annual charges on an investment of \$150 per kw might be reduced to 6%. Fuel cost is at 30 cents per million Btu. Total cost of power can be conveniently expressed as:

$$\begin{aligned} \text{\$ per kw per yr} &= K_1 + K_2 + 8760 K_3 \times \text{capacity factor} \\ \text{where } K_1 &= \text{capacity charge, \$ per yr} \\ K_2 &= \text{peak prepared charge, \$ per yr} \\ K_3 &= \text{energy cost, mills per kw-hr} \end{aligned}$$

Costs of representative power plants are summarized in Table 4. [T.B.]

**Bibliography:** E. Ayres and C. A. Scarlott, *Energy Sources—the Wealth of the World*, 1952; Babcock and Wilcox Company, *Steam, Its Generation and Use*; H. K. Barrows, *Water Power Engineering*, 3d ed., 1943; T. Baumeister (ed.), *Marks' Mechanical Engineers' Handbook*, 6th ed., 1958; C. F. Bonilla, *Nuclear Engineering*, 1957; B. G. A. Skrotzki and W. A. Vopat, *Applied Energy Conversion*, 1945.

## Power supply, electronic

A source of electric energy employed to furnish the tubes and semiconductor devices of an electronic circuit with the proper electric voltages and currents for their operation. The more common sources of energy are chemical batteries and alternating-current mains or lines. Batteries are useful as portable sources but are expensive and have small capacities. Alternating-current mains are not portable but are relatively inexpensive and have a large capacity.

One common method of classifying power supplies is by their use in electronic circuits. Most vacuum and gas tubes require a source of energy to energize their filaments or heaters; this type is known as a filament or heater power supply, or simply as an A supply. Tubes require plate or anode voltages, and transistors need voltages for their collectors and other elements. These voltages are supplied by a plate or anode power supply, sometimes called a B supply. Similarly the tubes may also require a control-grid power supply, often called a C supply, and additional power supplies for the screen grid and suppressor grid.

If the source is one of alternating voltage, a transformer is usually used to raise or lower the voltage to the required level. The alternating current (ac) usually must be changed to a direct current (dc); this is accomplished by a rectifier. A rectifier allows current to flow mostly in one direction, and a pulsating current results. This pulsating direct current is not suitable for most purposes and must be smoothed by a power-supply filter or voltage regulator.

A power-supply filter stores energy when the current is high and gives it up when the current is lower. The net result is to smooth out the variations in the current. The voltage regulator performs a similar function, but its operation is quite different. The gas-tube type of voltage regulator has a characteristic constant voltage over a large range of current values. Vacuum-tube regulators usually operate as variable resistances; the resistance decreases when the load is heavy and increases when the load is lighter. Voltage regulators deliver an almost constant output voltage in spite of variations in the load or in the input ac supply voltage. See VOLTAGE REGULATOR.

**Filament or heater power supply.** The filament or heater power supply is used to heat the filaments of vacuum or gas tubes so that electrons may be emitted from the filaments. If an indirectly heated cathode is used, a heater element inside the cathode must heat the cathode to a temperature at which electrons will be emitted.

Most of the present-day tubes heated by alternating current use 6.3- or 12.6-volt heaters. A step-down transformer is employed to change the 115 volts of the ac lines to these voltages. As an example, the heaters of four tubes may be connected in parallel across the secondary of a heater transformer whose primary is connected to the 115-volt ac lines. The center tap or one side of the secondary is often grounded. The tubes might take 0.5, 0.5, 1.0, and 2.0 amperes, respectively, for a total of 4.0 amperes. The heater transformer must have a rated capacity of at least 4.0 amperes to supply these tubes. The voltage that the insulation must withstand is also often stipulated. Some tubes require other heater voltages, such as 2.5, 5, and 25 volts. Therefore, heater transformers with more than one secondary are useful.

Heater transformers are both heavy and expensive. To eliminate transformers, heaters may be con-

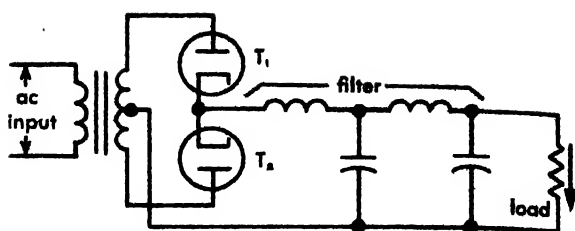


Fig. 1. Typical plate power supply.

nected in series across the ac line. Here each heater must take the same current, and the sum of the heater voltage ratings must be nearly equal to the ac line voltage. Other tubes use heaters made for direct connection across the ac lines.

For portable operation, batteries are usually used to heat the filaments. The tubes usually require 1-3 volts, so that one or two cells of a dry battery or one cell of a wet battery will supply sufficient voltage. Tubes having 6.3- or 12.6-volt heaters may be operated from a wet storage battery, such as an automobile battery, but the weight of the battery decreases its portability. The filaments of the tubes are connected in parallel and one side of the battery is connected to the plate supply, so a complete circuit exists for the electrons flowing in the tube.

**Plate power supply.** The plate power supply usually supplies direct current of about 50 to 300 or 400 volts, depending on the tube and the application. Figure 1 shows the schematic diagram of a typical supply. The ac line energizes the primary of the transformer; the secondary is connected to the anodes of the rectifier tubes  $T_1$  and  $T_2$ . The heater connections of the rectifier tubes are omitted for clarity. The common cathode connection is connected to the double-section  $L$  filter. This is the positive side of the filter. The negative side, usually grounded, is connected to the center tap of the transformer secondary. The load resistance is shown at the other end of the filter. A bleeder resistor is sometimes connected across the load to prevent the output voltage from increasing to dangerous values when the load is light.

The power, or plate, transformer should be selected to have a sufficiently high secondary voltage to supply the desired output voltage. Allowance must be made for the voltage drop or rise caused by the type of rectifier circuit employed, and for the voltage drop in the rectifier tubes and the filter elements caused by the combined load and bleeder currents. Also, the transformer secondary current rating must be sufficient to supply the combined currents. The rectifier tube or tubes must have voltage ratings sufficiently large for the transformer secondary voltages and current ratings sufficiently large for the combined load and bleeder currents and for the peak currents encountered in the particular rectifier circuit used. The filter capacitors should have sufficient capacitance for smoothing the ripple of the output voltage and sufficient voltage rating. The filter inductors should have sufficient inductance for smoothing, low resistance to the current carried, and high insulation voltage.

As the line-voltage fluctuates, the output dc voltage will also fluctuate. One method of reducing this fluctuation is to regulate the voltage of the ac lines by static ac voltage regulators.

**Control-grid power supply.** The control-grid power supply is used to supply voltages of 1–100 volts grid bias to the control grid of a vacuum tube. This is usually done by the use of a biasing resistor in series with the B supply and the vacuum tube, but a separate C supply similar to the above-mentioned B supply is sometimes employed. Usually the voltage and current requirements are much smaller for the C supply than for the B supply. Similar supplies are used occasionally to furnish the voltages required for the screen grids and suppressor grids of vacuum tubes.

**High-voltage power supply.** High-voltage power supplies are employed to supply dc voltages of 1–20 kilovolts or more, usually at currents of a few milliamperes or less. Voltage-doubling and voltage-multiplying rectifier circuits are useful in such applications, which include the cathode-ray tube power supplies of television and radar (see VOLTAGE-MULTIPLIER CIRCUIT). Another method of obtaining high voltages is to use a fly-back power supply circuit with a half-wave rectifier and filter or with a voltage-multiplier circuit. For heavy-current high-voltage supplies, circuits such as that of Fig. 1 are almost always used. All components of the circuit must then be selected for the requirements of the high voltage and high currents.

**Battery power supplies.** These are useful for portable applications where lower voltages and lower currents are usual. This is particularly true for transistor circuits. Dry battery units are used for A, B, and C supply requirements. Wet storage batteries may be used directly for the heaters of tubes and may also supply B and C power requirements by the use of vibrator or dynamotor power supplies.

**Power-supply filters.** Power-supply filters are usually of the low-pass LC variety, with the cut-off frequency of the filter made as much below the fundamental frequency of the ripple voltage as is economically feasible. Occasionally, low-pass RC filters will be used if the load current requirements are small. Generally lower cut-off frequencies require larger inductances and capacitances.

There are two subdivisions of these filters. The first, called an inductive-input filter, has a series inductor immediately after the rectifier. The second is the capacitive-input filter, which has a shunt capacitor immediately after the rectifier. One or more sections may be used, as shown in Fig. 2. Figure 2a shows a single shunt capacitor as the smoothing element. Figure 2b shows a two-section inductive-input filter. Figure 2c gives a two-section capacitive-input filter. The inductive-input filter has the disadvantage of giving a lower output voltage but the advantage of a better voltage regulation than the capacitive-input filter. The capacitive-input filter has the important disadvantage of producing a much higher peak current in the rectifiers.

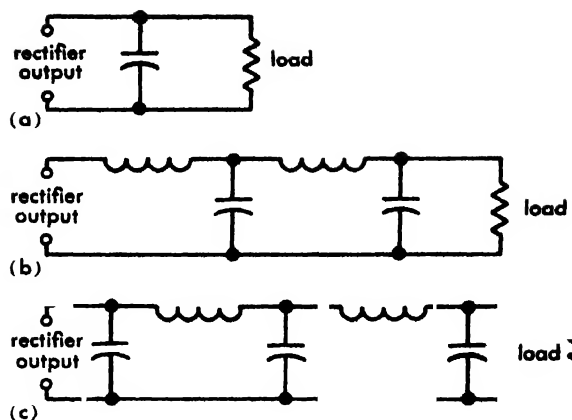


Fig. 2. Smoothing filters. (a) Shunt capacitor. (b) Inductive-input filter. (c) Capacitive-input filter.

The inductors, sometimes called choke coils, almost always have iron cores, and the resistances of the windings are kept low for high efficiency and good voltage regulation. Iron-cored inductors can be made to have an inductance that increases considerably as the load current approaches zero. These inductors, called swinging choke coils, help prevent the large rise in voltage at low load currents in capacitive-input filters. Capacitors for voltages below 500 volts are almost always electrolytic capacitors, but above that voltage they have paper or plastic insulation. See CAPACITOR; INDUCTOR; for other types and general treatment of electronic circuits, see CIRCUIT, ELECTRONIC.

[D.L.W.]

**Bibliography:** J. M. Carroll, *Transistor Circuits and Applications*, 1957; H. E. Clifford and A. H. Wing (eds.), *Electronic Circuits and Tubes*, 1947; J. Millman and S. Seely, *Electronics*, 2d ed., 1951; H. J. Reich, *Theory and Applications of Electron Tubes*, 2d ed., 1944; S. Seely, *Electronic Engineering*, 1956.

## Power-factor meter

An instrument used to indicate whether load currents and voltages are in time-phase with one another.

In an alternating-current circuit, the current  $I$  may lead or lag the voltage  $E$  by some angle  $\phi$ , and the average power  $P$  would be the average of

$$(E_{\max} \cos \omega t) [I_{\max} \cos (\omega t \pm \phi)]$$

or

$$P = EI \cos \phi$$

The factor  $\cos \phi = P/EI$  is called the power factor of the load. See ALTERNATING-CURRENT CIRCUIT THEORY.

The formula  $P = EI \cos \phi$  shows that if two customers use equal currents the energy loss in the line is the same for the two. If one of them has a lower power factor he uses less energy, and a rate based only on energy delivered would involve an unequal charge for service.

**Single-phase power-factor meter.** This instrument contains a fixed coil that carries the load current, and crossed coils that are connected to the

load voltage. There is no spring to restrain the moving system, which takes a position to indicate the angle between the current and voltage. The scale can be marked in degrees or in power factor. For a complete discussion of this instrument see PHASE METER.

The angle between the currents in the crossed coils is a function of frequency, and consequently each power-factor meter is designed for a single frequency and will be in error at all other frequencies. For low harmonic content the indication of the power-factor meter is usually accepted for contract purposes.

**Polyphase power-factor meters.** These meters are usually designed differently from single-phase meters. For balanced polyphase loads their application is straightforward. For a quarter-phase load, two voltages 90° out of phase are present and each can supply one of the crossed coils of the power-factor meter. If the voltages are unequal or their relative phase angle is different from 90°, the indication will no longer have meaning.

Three-phase power-factor meters are built to connect the fixed coil in one line and the crossed coils between the other two lines. The meter is constructed with a 60° angle between the crossed coils. An unbalance in either voltage or angle will cause erroneous readings.

In the case of balanced four-phase and balanced six-phase loads, single-phase power-factor meters can be used directly without any correction.

In the case of unbalanced polyphase circuits with harmonics, various portions of the load may have different power factors and for the combined load there is no common phase angle; consequently the statement that the power factor is equal to  $\cos \phi$  becomes meaningless. See WAVEFORM, NONSINUSOIDAL. [H.S.O.]

**Bibliography:** F. A. Laws, *Electrical Measurements*, 2d ed., 1938; M. B. Stout, *Basic Electrical Measurements*, 1950.

## Poynting's vector

A vector, the outward normal component of which, when integrated over a closed surface in an electromagnetic field, represents the outward flow of energy through that surface. It is given by the equation

$$\Pi = \mathbf{E} \times \mathbf{H} = \mu^{-1} \mathbf{E} \times \mathbf{B} \quad (1)$$

where  $\mathbf{E}$  is the electric field strength,  $\mathbf{H}$  the magnetic field strength,  $\mathbf{B}$  the magnetic flux density, and  $\mu$  the permeability. This can be shown with the aid of Maxwell's equations:

$$\begin{aligned} \mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H}) &= \nabla \cdot (\mathbf{E} \times \mathbf{H}) \\ &= -\mathbf{i} \cdot \mathbf{E} - \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} - \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} \end{aligned} \quad (2)$$

where  $\mathbf{D}$  is the electric displacement and  $\mathbf{i}$  the current density. Integration over any volume  $v$  and use of the divergence theorem to replace one volume integral by a surface integral give

$$-\int (\mathbf{E} \times \mathbf{H}) \cdot \mathbf{n} dS$$

$$\int_v \left[ \frac{\partial}{\partial t} (\tfrac{1}{2} \mathbf{B} \cdot \mathbf{H}) + \frac{\partial}{\partial t} (\tfrac{1}{2} \mathbf{D} \cdot \mathbf{E}) + \mathbf{E} \cdot \mathbf{i} \right] dv \quad (3)$$

where  $\mathbf{n}$  is a unit vector normal to  $dS$ . In the volume integral,  $\tfrac{1}{2} \mathbf{B} \cdot \mathbf{H}$  is the magnetostatic energy density and  $\tfrac{1}{2} \mathbf{D} \cdot \mathbf{E}$  is the electrostatic energy density, so the integral of the first two terms represents the rate of increase of energy stored in the magnetic and electric fields in  $v$ . The product of  $\mathbf{E} \cdot \mathbf{i}$  is the rate of energy dissipation per unit volume as heat or, if there is a motion of free charges so that  $\mathbf{i}$  is replaced by  $\rho \mathbf{v}$ ,  $\rho$  being the charge density, it is the energy per unit volume used in accelerating these charges. The net energy change must be supplied through the surface, which explains the interpretation of Poynting's vector.

It should be noted that this proof permits an interpretation of Poynting's vector only when it is integrated over a closed surface. In quantum theory, where the photons are localized, it could be interpreted as representing the statistical distribution of photons over the surface. Perhaps this justifies the common practice of using Poynting's vector to calculate the energy flow through a portion of a surface.

When an electromagnetic wave is incident on a conducting or absorbing surface, theory predicts that it should exert a force on the surface in the direction of Poynting's vector. See RADIATION PRESSURE; see also ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; WAVE EQUATION. [W.R.S.M.]

## Prairie

A land of tall grasses occurring in the humid or subhumid parts of the mid-latitudes between the forests and the semiarid short-grass steppes. The North American prairies are better known than those of other continents. The tall-grass prairies occur on the more moist side toward the forest margins. They consist of tall grasses, 5-8 ft in height, together with a variety of flowering plants;

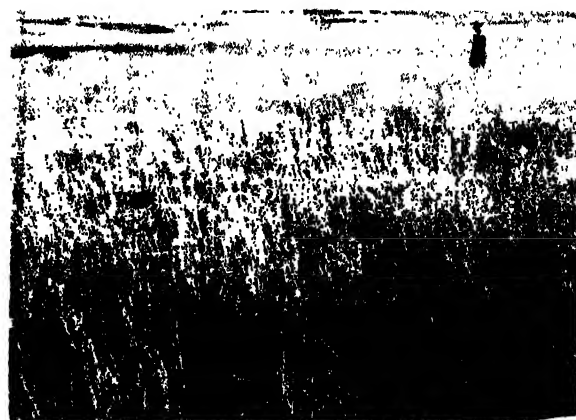


Fig. 1. Mixed-grass prairie near Agate, Nebraska. The short-grass parts are a foot or less in height. (From J. E. Weaver and F. E. Clements, *Plant Ecology*, 2d ed., McGraw-Hill, 1938)

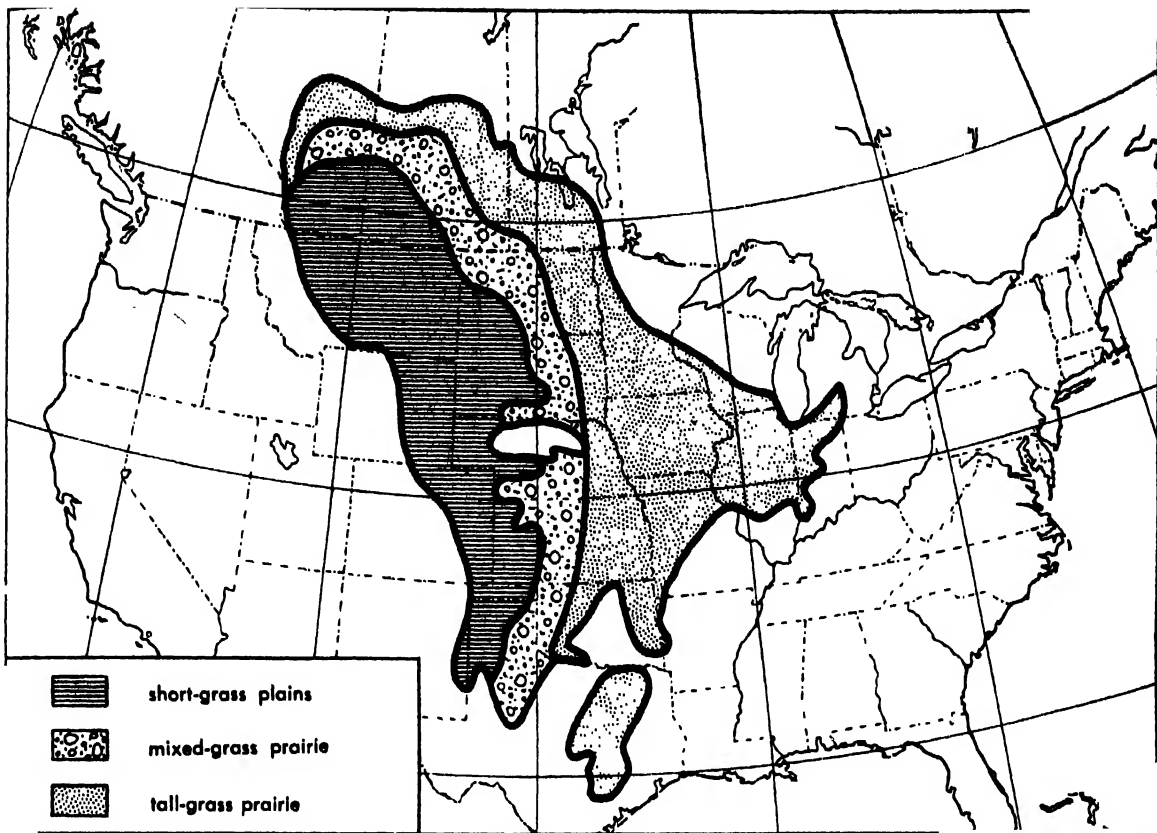


Fig. 2. Types of natural grasslands east of the Rocky Mountains. The short grass covers the Great Plains and the mixed grass is transitional to the tall-grass prairie.

(After J. Richard Carpenter in V. C. Finch, G. I. Trewartha, E. H. Hammond, and A. H. Robinson, *Elements of Geography*, 4th ed., McGraw-Hill, 1957)

trees are absent or rare on the uplands. Toward the dry side the mixed-grass prairies are apparently a transition into the short-grass steppes. They bear some grasses 2-4 ft tall and others which are less than 1 ft high. Prairies grow in climatic situations where there is enough rainfall for crops. They occur on relatively level land, and have developed deep, dark, fertile soils. For these reasons most prairie vegetation has been displaced by agriculture and little original prairie land remains in the world. The prairies of North America formerly bore an extensive animal population of birds and small mammals together with grazing animals, chiefly the bison. These animals disappeared with the conversion of the prairies into cropland. See STEPPE. [C.M.D.]

## Prairie chicken

Either of two American game birds of the genus *Tympanuchus*, family Tetraonidae, the grouse family. The greater prairie chicken, *T. cupido*, formerly was an important game bird of the Great Plains, but it has disappeared entirely throughout much of its original range and is rare in the remainder. With changes in agriculture, the principal population of this bird has shifted into the Canadian prairie region in recent years. The eastern representative of this species, the heath hen,



The prairie chicken, *Tympanuchus cupido*; length to 18½ in. (Alfred M. Bailey, *National Audubon Society*)

formerly found on the Atlantic coastal plain, is now extinct. The lesser prairie chicken, *T. pallidicinctus*, is paler than *T. cupido*, and is resident in the southwestern plains area.



**Prairie chickens** are noted for their elaborate mating rituals, involving participation in dancing and drumming (booming) by aggregations of males on selected booming grounds. See GALLIFORMES: GROUSE. [J.D.S.]

## Prairie dog

Any of three species comprising the genus *Cynomys*. Prairie dogs are rather large, stout-bodied, terrestrial squirrels, with small ears and short, flat tails. They live in large colonies, called towns, digging burrows for their homes, each surrounded by a mound of earth. Of the two species in the United



**The prairie dog, *Cynomys ludovicianus*; length to 14½ in. (Grace A. Thompson, National Audubon Society)**

\*States one hibernates, and the other is active throughout the winter. They feed mainly upon plant material but will also eat insects. They are much less common than originally, many of their great towns now being completely wiped out. In 1900 one colony in Texas was estimated to be 100-150 miles wide and 250 miles long. Burrowing owls, rattlesnakes, and the rare black-footed ferret prey upon young prairie dogs and are unwelcome predators in their burrows. See RODENTIA. [J.D.B.]

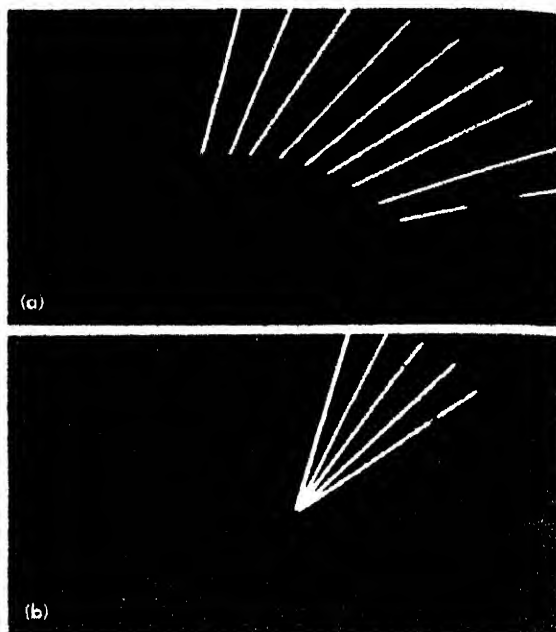
## Prandtl-Meyer expansion fan

A steady planar compressible fluid flow that occurs only at supersonic speeds. The Prandtl-Meyer expansion fan is essentially the isentropic flow around a corner from a uniform supersonic flow. The illustration shows a typical expansion fan, starting from the uniform flow at Mach number  $M_1$  along a flat wall, and turning a corner through an angle  $\theta$ . The straight lines in the fan are Mach waves, which are standing sound waves in the fluid. The flow behind the fan is also uniform, but the velocity is higher and the pressure is lower. A centered Prandtl-Meyer expansion fan develops at a sharp corner in a supersonic flow. All flow quantities are constant on a given Mach wave in the fan and depend only on  $M_1$  and the angle through which the flow has turned. Using  $M_1 = 1$  as a convenient reference, the relationship between flow angle  $\theta$  and local Mach number  $M$  is

$$\theta = \sqrt{\frac{\gamma+1}{\gamma-1}} \arctan \sqrt{\frac{\gamma-1}{\gamma+1}} \sqrt{M^2-1} - \arctan \sqrt{M^2-1}$$

The local pressure  $p$  is obtained from

$$\frac{p_0}{p} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{\gamma/(\gamma-1)}$$

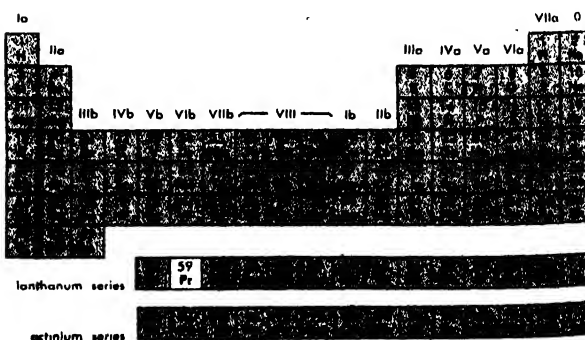


**Supersonic flow around a curve (a) produces a fan of standing sound waves. At a sharp corner (b), the expansion fan is centered.**

where  $p_0$  is the upstream total pressure, and  $\gamma = C_p/C_v$ , where  $C_p$  = specific heat at constant pressure, and  $C_v$  = specific heat at constant volume. See FLUID-FLOW PRINCIPLES. [A.E.BR.]

## Praseodymium

A chemical element, Pr, atomic number 59, and atomic weight 140.91. Praseodymium is a metallic element of the rare-earth group. The stable isotope 140.91 makes up 100% of the naturally oc-



curing element. It was discovered by C. F. Auer von Welsbach in 1885 when he separated the salts of the so-called element didymium into two fractions, praseodymium and neodymium. The oxide is a black powder, the composition of which varies according to the method of preparation. It is usually considered to be  $\text{Pr}_2\text{O}_{11}$ , although if oxidized under a high pressure of oxygen it can approach the composition  $\text{PrO}_2$ . It can be reduced in hydrogen to give a pale green  $\text{Pr}_2\text{O}_3$ . The oxide  $\text{PrO}_2$  is the only form in which it has been clearly demonstrated that praseodymium exists in the quadrivalent

lent form. The black oxide dissolves in acid with the liberation of oxygen to give green solutions or green salts. For properties of the metal, see RARE-EARTH ELEMENTS.

The salts have found application in the ceramic industry for coloring glass and for glazes.

[F.H.SP.]

## Preamplifier

A voltage amplifier suitable for operation with a low-level input signal. It is intended to be connected to another amplifier with a higher input level. Preamplifiers are necessary when an audio amplifier is to be used with low-output transducers such as magnetic phonograph pickups. The preamplifier may also incorporate frequency-correcting networks to compensate for the frequency characteristics of a given transducer and make the frequency response of the preamplifier-amplifier combination uniform. See VOLTAGE AMPLIFIER.

The design of preamplifiers is critical, because the input signal level is low and the amplification is high. The hum introduced by tubes and the noise voltages from resistors and vacuum tubes must be closely controlled. Furthermore, the preamplifier must be shielded from external magnetic fields which would induce a voltage in the circuit. An additional source of undesirable voltages is microphonics. For a discussion of these undesirable conditions see AMPLIFIER.

[H.F.K.]

## Precambrian

The name for rocks older than the Cambrian or earliest period in the earth's history in which abundant fossils indicate the existence of life. Because



of their older age, Precambrian rocks lie at the base of the succession of formations, and, on the whole, have undergone more transformation than the rocks of later age. In places, Cambrian strata pass transitionally downward into the Precambrian, but more commonly they are separated by an interval of erosion during which the rocks were worn down to produce an unconformity. See CAMBRIAN; UNCONFORMITY.

Fossils, so helpful in determining the age of younger rocks, are almost entirely absent in the Precambrian. Therefore, the sequence of rocks within the Precambrian system must be deter-

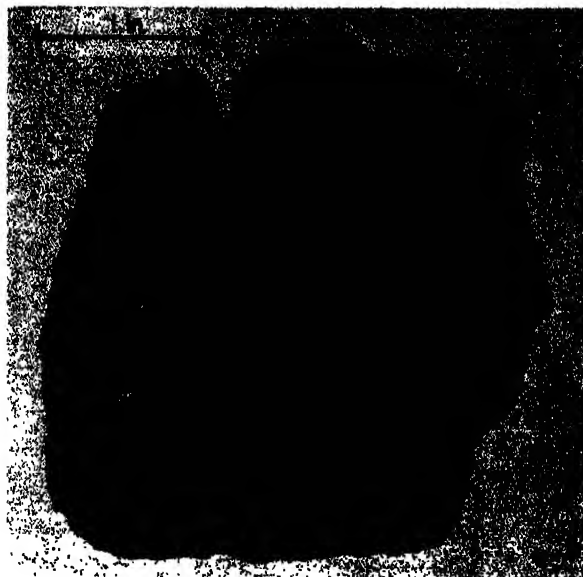


Fig. 1. Striated and faceted pebble from Precambrian conglomerate of glacial origin in the Kekeko Hills, Western Quebec, Canada. (M. E. Wilson, Geol. Survey, Canada)

mined by detailed structural examination, by the relationships to unconformities or igneous intrusions, or from the ratios of isotopes in radioactive minerals determined by geophysical methods. Radioactive minerals, however, commonly occur in folded, mountain-built rocks, so that their age is more nearly that of the mountain building than the age of the enclosing rock, which may be millions of years older. The maximum age of Precambrian minerals, so far determined by geophysical methods, is about  $3 \times 10^9$  years. See RADIOACTIVE MINERALS; ROCK (AGE DETERMINATION).

**Occurrence.** Precambrian rocks are widely distributed. They occur on all the continents and underlie about one-fifth of the earth's surface. Structurally, they occur mainly in two types of areas: (1) in the interior of partly denuded mountains; (2) in extensive areas of relatively low elevation called shields. When mountain chains are formed by folding and uplift, Precambrian rocks commonly intruded by masses of granite or related igneous rocks are thrust upward by the mountain-building movement. However, as erosion of the mountains continues, the Precambrian rocks and their associated igneous intrusions become exposed. The shield areas of the earth also are underlain mainly by mountain-built rocks, but have been so denuded that they are mountains structurally and not by elevation.

The oldest Precambrian surficial rocks are chiefly lavas, fragmental material blown explosively from volcanoes, and bedded sediments. The sediments are notably gravels (conglomerate), sandstone or its transformed equivalent quartzite, mudlike deposits (graywacke), or less commonly limestone. These rocks are similar to those being formed on the earth's surface today. Furthermore, in some areas

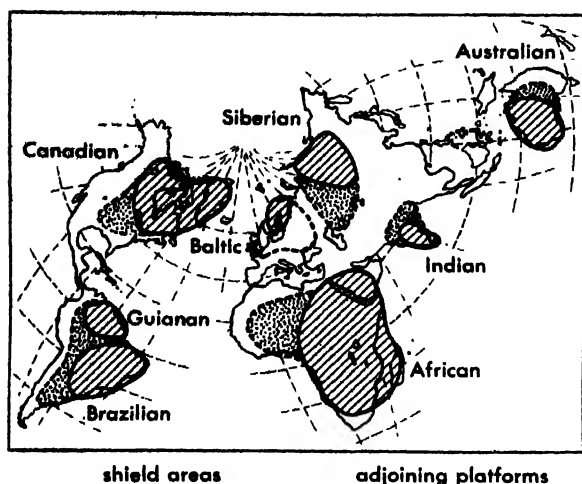


Fig. 2. Shield areas and adjoining platforms composed of Precambrian rocks. (From R. C. Moore, *Introduction to Historical Geology*, 2d ed., McGraw-Hill, 1958)

conglomerate is found containing striated and faceted pebbles, cobbles, or boulders up to 5 or even 10 ft in diameter, enclosed in a massive unstratified matrix. Such deposits, characteristically derived from glaciers, indicate that the Precambrian climate, at least at times, was cold (Fig. 1). All of these observations provide evidence that, if the earth was originally molten, as most hypotheses of its origin assume, all its original crust must have been destroyed before the existing rocks were formed.

**Shield areas.** Large relatively stable areas with broad relatively flat surfaces, or low topographic relief, are characteristic of many Precambrian regions. These areas stood firm or moved differentially upward in contrast to areas of subsidence (Fig. 2).

**Canada.** The Canadian Shield, the world's largest shield area, occupies most of northeastern Canada, Greenland, the Adirondack region of New York, and the northern parts of Michigan, Wisconsin, and Minnesota. Most of this immense area has been mapped geologically in an approximate lithological way, but the complete succession of formations has been determined only in widely separate local areas. For this reason a tentative classification of the rocks of the shield into two major parts, the Early Precambrian or Archean, and the Late Precambrian or Proterozoic, is, in the present state of knowledge, the only subdivision practicable.

The structural succession of formations has been determined locally in the shield by detailed examination, using the criteria of top determination of sedimentary beds and lava flows. The most important of these criteria are graded bedding in graywacke and pillow tops in lavas. Graded bedding forms in graywacke by the more rapid settlement of coarse particles than fine particles during deposition (Fig. 3). Pillows are believed to originate by the globulation of lava where extruded into water. The slow descent of the globules per-

mits the formation of a rounded pillow top crust before succeeding globules are deposited (Fig. 4). These criteria of top determination, combined with the relationships of lavas and sediments to intrusions of igneous rocks, make it possible to determine both the succession of formations and their division into separate groups or series by unconformities throughout areas of considerable extent. However, the correlation of rocks across the wide geographical or geological barriers between these areas is less certain. It has become the custom, therefore, to use local names of formations in areas where geological mapping has been most detailed.

The surficial rocks of the Canadian Shield belong mainly to four classes: (1) widely distributed, basal lava flows and poorly sorted sediments of the type commonly deposited in the sea adjacent to mountains; (2) crystalline limestone, pure quartzite, shale or slate, completely sorted rocks of the types laid down in the sea adjacent to deeply weathered land of low relief; (3) shale, slate, iron-bearing formation, limestone and dolomite, sediments probably deposited in elongated basins (in places the limestone and dolomite contain concentric structures believed to be of algal origin); (4) interbedded conglomerate, sandstone, and lava flows laid down on land. The rocks of class 1 have been intensely folded and intruded by granite. They belong to the Archean. The strata of classes 3 and 4 lie at the top of the Precambrian succession and have been folded only in places. They belong, therefore, to the Proterozoic. The sedimentary rocks of class 2 have all been mountain-built and intruded by granite, but the folding and recrystallization

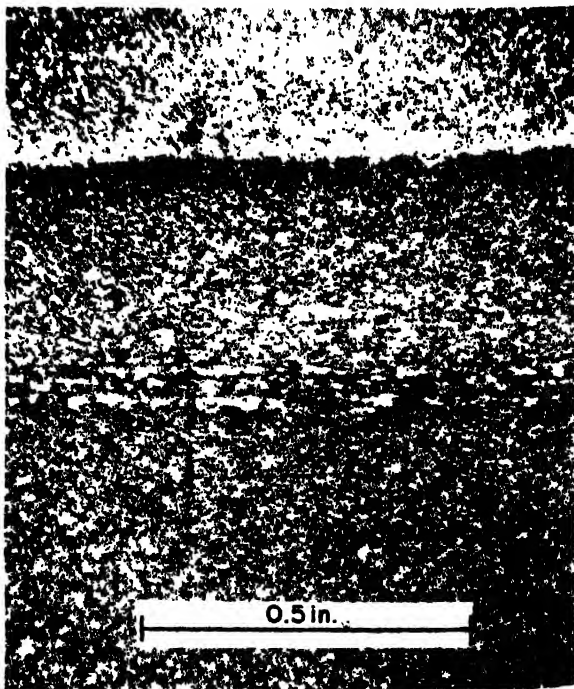


Fig. 3. Microphotograph of graded bedding in Timiskaming graywacke, Rouyn Township, Western Quebec, Canada. (M. E. Wilson, Geol. Survey, Canada)

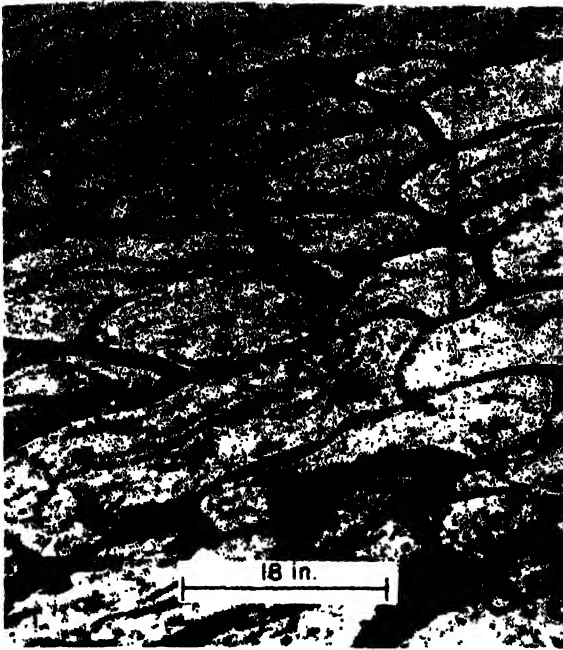


Fig. 4. Pillow structure in lavas of Yellowknife Group, Northwest Territories, Canada. Top of flow towards upper left hand corner. Length of hammer 18 in. (J. F. Henderson, Geol. Survey, Canada)

vary greatly in intensity. Part of them is classed as Archean and part as Proterozoic.

**The Baltic.** The Baltic Shield, the only Precambrian area of the shield class in Europe, occupies part of Norway, most of Sweden, Finland, and the Kola peninsula of U.S.S.R. Its rocks are divided by unconformities into three or possibly four groups. The oldest resemble the basal formations of the Canadian Shield and are called Archean. The youngest are similar to the Keweenaw sandstones and lavas of the Lake Superior region and are classed as Proterozoic.

**Asia.** The Precambrian shield areas of Asia are in Siberia and the peninsula of India. The shields of Siberia are known as the Anabar and Aldan areas. They are composed of folded Early Precambrian sediments and granite surrounded by formations of Late Precambrian age. In the peninsula of India the Precambrian rocks, because of their similarity to the early rocks of the Canadian Shield, are called Archean, and the younger rocks are called Purana. The formations of the Purana group are separated from the Archean or Dharwar system by a major unconformity and have been only moderately deformed.

**Australia.** The Australian Shield occupies most of western and northwestern Australia. Its formations occur in three major groups, the older two of which are classed as Archean and the third as Proterozoic. The surprising resemblance of the rocks of the Australian Central Goldfields region to the Precambrian of the Canadian Shield has been noted by H. E. McKinstry.

**Africa.** The areas of Precambrian rocks in Africa that have been called shields occur irregularly from

the Suez region southward to the Transvaal. In all of this territory, rocks called Basement, Archean, or Primitive systems are present. Late Precambrian rocks to which local names have been given are also represented. In the Transvaal, Precambrian rocks of Proterozoic age have been intruded by an assemblage of igneous rocks called the Bushveld complex.

**South America.** The Precambrian rocks of South America occur mainly in two extensive areas known as the Brazilian and Guianian Shields. The basal rocks of the Brazilian Shield are largely crystalline limestone and schist intruded by granite; in the Guianian Shield they consist of lavas, schists, gneisses, and granite. The Late Precambrian rocks in both shields consist of quartzite, iron formation, and other unfossiliferous sediments.

**Mountain areas.** Precambrian rocks occur locally in many of the world's mountains. This type of occurrence is well exemplified in the cordilleran region of North America and the Highlands of Scotland. In the former locality, an outstanding cross section of the Precambrian is exposed in the Grand Canyon of the Colorado River, where Archean schists and granite are overlain with great unconformity by the Grand Canyon Series. The most widespread Precambrian rocks, however, are the Late Precambrian Beltian sediments which underlie extensive areas in the interior of British Columbia and most of northern Montana and Idaho. These rocks are divided by an unconformity into two series. They have a maximum thickness of over 60,000 ft.

In the Highlands of Scotland the succession of formations is complicated by post-Precambrian mountain building and faulting. In the northwest Highlands, Early Precambrian transformed sediments and igneous gneisses of the Lewisian Series are overlain unconformably by the little disturbed Torridonian conglomerate and sandstone presumed to be of Late Precambrian age. In the adjacent highlands to the south are the highly recrystallized sediments of the Moine and Dalradian Series.

[M.E.W.]

**Bibliography:** C. O. Dunbar, *Historical Geology*, 2d ed., 1959; H. E. McKinstry, *Precambrian problems in western Australia*, *Am. J. Sci.*, 243A: 448-466, 1945; T. W. E. David, *Geology of the Commonwealth of Australia*, vol. 1, 1950; M. E. Wilson, *Precambrian classification and correlation in the Canadian Shield*, *Bull. Geol. Soc. Am.*, 69(6): 757-774, 1958.

## Precast concrete

A concrete structure may be constructed by casting the concrete in place on the site, by building it of components cast elsewhere, or by a combination of the two. Concrete cast in other than its final position is called precast.

In contrast with cast-in-place concrete construction, in which columns, beams, girders, and slabs are cast integrally or bonded together by successive pours, precast concrete requires field connections

to tie the structure together. These connections can be a major design problem.

Form costs are much less with precast concrete because the forms do not have to be supported on falsework in the structure. They may be set on the ground in a convenient position. Furthermore, a thin wall is difficult to concrete if it must be cast vertically because the concrete has to be placed in the narrow opening at the top of the form. Such a wall is easily precast flat on the ground. Moreover, the large side forms are eliminated and so are the many braces needed to keep a vertical form in place.

With some types of precast concrete construction, no time is lost in waiting for concrete to gain strength at one level of a structure before the next level can be placed; such delays are common with cast-in-place construction. Frequently, the permanent precast units can be used as a working platform, eliminating the need for a temporary deck.

Precast units can be standardized. Savings can then result from repeated reuse of forms and assembly-line production. Furthermore, high quality can be maintained because of the controls that can be kept on production under plant conditions. However, there is always the possibility that transportation, handling, and erection costs for the precast units will offset the savings. See CONCRETE; CONCRETE SLAB.

**Floor and roof systems.** Precast concrete floor and roof systems may be similar to what is generally used for cast-in-place construction. The components may be bolted together, seated on each other or on brackets, held by friction devices, prestressed, or the reinforcing of adjoining members welded and the gap between them filled with cast-in-place concrete. Also, certain systems peculiar to precast construction may be used:

1. I-beam type with cast-in-place or precast slab.
2. Hollow-core-type joists.
3. Assembled concrete-block type
  - a. With contact faces between units ground to provide a slight camber to the assembly.
  - b. With contact faces parallel but with a tension in the lower moment bars sufficient to align and hold the assembly together, and to provide a slight camber.
4. Precast inverted T-beam joists with precast fillers between.
5. Integrally precast slab and T-beam joists.

**Tilt-up construction.** Originally, tilt-up construction was the name given to a method of precasting walls in which the units were cast on the ground at the place where they were to be erected, then tilted up to the vertical and anchored when they had gained sufficient strength. Later it became customary to refer to all types of precast wall construction as tilt-up construction.

Generally, the wall is concreted on a casting platform. Only side forms are needed. Sometimes it is advantageous, not only for walls but for floor and roof panels as well, to cast successive units one on top of the other. A bond-breaking agent is applied

to the surface of the casting platform and between successive units.

Inserts usually are cast in the panels to facilitate lifting. The precast units may be lifted with a crane or A frame, often equipped with a strong-back, or frame, to distribute the uplift forces evenly.

**Lift-slab or Youtz-Slick method.** In one type of precast construction for buildings, popularly known as lift-slab but sometimes called Youtz-Slick after the developers of the method, floor and roof slabs for a multistory building are cast on the ground around the columns. The slabs are cast one on top of the other, with a bond-breaking agent between them. Jacks atop the columns lift them to their final position, where they are anchored. [F.S.M.]

## Precession

An angular velocity of the axis of spin of a spinning rigid body, which arises as a result of external torques acting on the body. Examples are the precession of a spinning top, the earth (precession of the equinoxes), an airplane propeller, or a gyroscope. Of these, the most familiar is undoubtedly the precession of the equinoxes, a slow change in the direction of orientation of the earth's axis of rotation which results in a gradual westward motion of the equinoxes. For a discussion of this phenomenon, which was known to the ancients, see PRECESSION OF EQUINOXES. The uniform precession of a charged spinning body in a uniform magnetic field is called Larmor precession. This motion, similar to that of a rapidly spinning top, is of great importance in atomic physics. See LARMOR PRECESSION.

**Motion of a spinning top.** Precession is best explained by a discussion of the behavior of a symmetrical spinning top or of any rigid body spinning about an axis of symmetry. Consider a top spinning rapidly about its  $z$  axis of symmetry and placed horizontally on a point support at  $O$  (Fig. 1). The top will not fall as a result of the pull of gravity (its weight,  $W$ ) as might be supposed, but rather one will observe that the  $z$  axis of the top rotates slowly about the vertical  $y$  axis while maintaining its position in the horizontal  $xz$  plane. This rotation

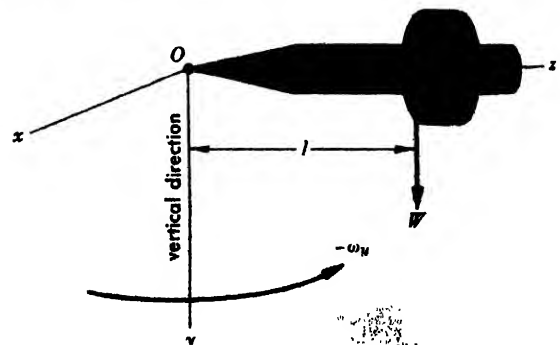


Fig. 1. A fast-spinning top supported at point  $O$  and released in a horizontal plane precesses about the vertical  $y$  axis as shown.

about the  $y$  axis is called precession and the motion can be predicted from the equation

$$\begin{aligned} \mathbf{M}_O = & \mathbf{i}(\dot{H}_x + \omega_y H_z - \omega_z H_y) \\ & + \mathbf{j}(\dot{H}_y + \omega_z H_x - \omega_x H_z) \\ & + \mathbf{k}(\dot{H}_z + \omega_x H_y - \omega_y H_x) \end{aligned} \quad (1)$$

where  $\mathbf{M}_O$  is the moment, or torque, about the point  $O$ ;  $H_x$ ,  $H_y$ , and  $H_z$  are the  $x$ ,  $y$ , and  $z$  components of the angular momentum  $\mathbf{H}$  of the top; and  $\mathbf{i}$ ,  $\mathbf{j}$ , and  $\mathbf{k}$  are unit vectors along  $x$ ,  $y$ , and  $z$  respectively. For the derivation of Eq. (1), see RIGID-BODY DYNAMICS.

If  $\mathbf{k}S$  is the spin velocity of the top about the  $z$  axis relative to the  $x$ ,  $y$ ,  $z$  coordinate system, then  $\omega_x$ ,  $\omega_y$ , and  $\omega_z$  are the angular velocity components of the  $x$ ,  $y$ , and  $z$  axes which are attached to the top at  $O$  and move with it except for this relative spin.

The angular momentum of the top is

$$\mathbf{H} = \mathbf{i}I_x\omega_x + \mathbf{j}I_y\omega_y + \mathbf{k}I_z(\omega_z + S) \quad (1a)$$

where  $I_x = I_y$ ,  $I_z$  are the moments of inertia of the top about the  $x$ ,  $y$ ,  $z$  axes through  $O$ .

Substitution of this equation into Eq. (1) (for steady-state motion  $\dot{\omega}_x = \dot{\omega}_y = \dot{\omega}_z = \dot{S} = 0$ ) gives

$$\begin{aligned} \mathbf{M}_O = & \mathbf{i}[\omega_y(\omega_z + S)I_z - \omega_x\omega_y I_x] \\ & + \mathbf{j}[\omega_z\omega_x I_x - \omega_x(\omega_z + S)I_z] \\ & + \mathbf{k}[\omega_x\omega_y I_z - \omega_y\omega_x I_x] \end{aligned} \quad (2)$$

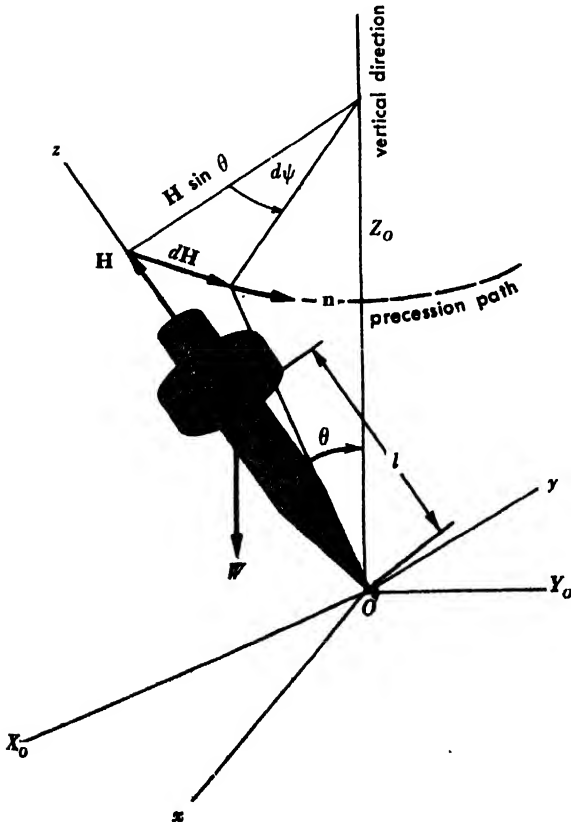


Fig. 2. A fast-spinning top supported at point  $O$  precesses about the vertical  $Z_0$  axis as shown.

For the case where  $\omega_x$ ,  $\omega_y$ , and  $\omega_z$  are negligible compared to the spin  $S$ , Eq. (2) becomes

$$M_x \cong S\omega_y I_z \quad (3a)$$

$$M_y \cong -S\omega_x I_z \quad (3b)$$

$$M_z = 0 \quad (3c)$$

For the case illustrated in Fig. 1,  $M_x = -Wl$ ;  $M_y = 0$  (no moment due to  $W$  about  $y$ ); and  $M_z = 0$  (no frictional torque about  $z$ ,  $\dot{S} = 0$ ).

From Eq. (3a), verified by experiment, the top slowly precesses about the  $y$  axis with a precessional velocity that is counterclockwise when observed from above and given by

$$\omega_y = \frac{-Wl}{I_z S} \quad (\text{a constant}) \quad (4)$$

From Eq. (3b),  $\omega_x = 0$ , and the top maintains its position in the horizontal plane. From Eqs. (3), one sees that any moment exerted on the spinning body about the  $x$  axis produces an angular velocity of precession about the  $y$  axis, and any moment exerted about the  $y$  axis produces a negative precession about the  $x$  axis. Such a spinning body is the heart of the gyroscope and such moments  $M_x$  and  $M_y$  are called gyroscopic moments.

If the top of Fig. 1 is now placed in the more general position shown in Fig. 2, a similar precession about a fixed vertical axis ( $Z_0$ ) due to the gravity moment will result. This is most easily seen by considering the angular momentum vector  $\mathbf{H}$ . Because  $S$  is very large as compared to the angular velocity components of the coordinate system  $x$ ,  $y$ ,  $z$  attached to the top, the vector  $\mathbf{H}$  is directed very nearly along the spin axis  $z$ , and is given by  $\mathbf{H} \cong \mathbf{k}I_z S$ . From Fig. 2, the moment vector due to gravity is directed normal to the plane of the  $z$  and  $Z_0$  axes at all times and this produces a change  $d\mathbf{H}$ , in time  $dt$ , of the angular momentum vector. From Fig. 2,  $d\mathbf{H} = H(\sin\theta)(d\psi)\mathbf{n}$  ( $\mathbf{n}$  = unit vector normal to plane  $zZ_0$ ),  $d\mathbf{H}/dt = H\sin\theta(d\psi/dt)\mathbf{n}$ , and  $\mathbf{M} = Wl(\sin\theta)\mathbf{n}$ . The equation of angular motion  $\mathbf{M} = d\mathbf{H}/dt$  becomes  $Wl\sin\theta = H\sin\theta(d\psi/dt)$ . Therefore

$$\frac{d\psi}{dt} \cong \frac{Wl}{I_z S} \quad (5)$$

is the precessional velocity.

The equation of motion is satisfied if the axis of the top swings or precesses around  $Z_0$  with a velocity  $Wl/I_z S$  and with  $\theta$ , the inclination of the top to the vertical, unchanged. Note that Eq. (5) is identical to Eq. (4) and for large spin the precession is small, satisfying the assumptions of the analysis, that the components of  $\omega$  are negligible compared to  $S$ .

When the value of  $S$  becomes smaller and no longer fulfills this inequality, the general motion of a spinning top must be considered. For the general case, the resultant motion consists of precession and a second angular velocity of the axis of spin called nutation. See NUTATION (ASTRONOMY AND MECHANICS).



**Gyroscope motion.** A gyroscope consists of a rapidly spinning rigid body with an axis of symmetry. Generally such a body is mounted in a Cardan's suspension (see POINSON'S METHOD) so that its mass center is fixed and set spinning with a large angular velocity. Equations (3) are valid and (3a) and (3b) can be expressed as a single vector equation

$$\omega = \frac{1}{I_z S^2} (\mathbf{S} \times \mathbf{M}) \quad (6)$$

where  $\mathbf{S}$  has a direction along  $z$ , the axis of spin. The precessional angular velocity  $\omega$  of the axis of spin is always at right angles to  $z$  and  $\mathbf{M}$ . Therefore, from Eq. (6), the axis of spin always turns toward the resultant moment acting on the gyroscope.

This simple analysis and result shows the stability which a large rotation imparts to a body, because under such rotation the body refuses to change the direction of its axis unless large moments are applied. When such moments are applied, the resulting displacements due to the precession are obtained according to Eq. (6) and are shown in Fig. 3. As a result a gyroscope is applicable for stabilizing ships in rolling seas or a monorail car, for inertial navigation instruments, and as the heart of the gyrocompass. Furthermore, the dynamical analysis of the gyroscope explains how bullets and shells that are spun obtain aerodynamic stability in flight. See BALLISTICS, EXTERIOR; GYROSCOPE.

**Airplane maneuvers.** Rotating masses on aircraft such as propellers, gas turbine rotors, jet engine compressors, and the like can exert gyroscopic moments on the airplane during maneuvers.

As an example, consider an airplane which is making a sharp turn in a horizontal plane (see Fig.

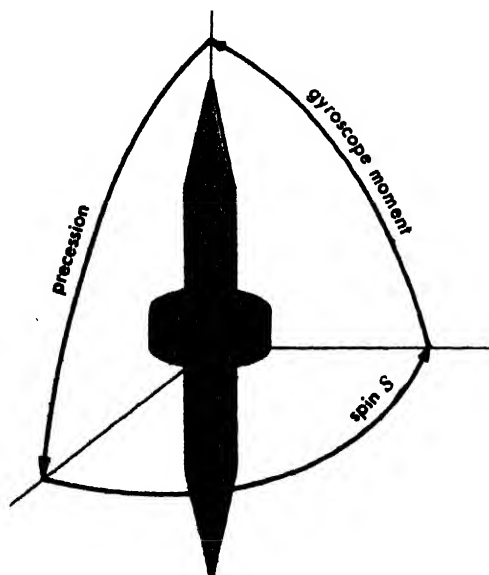


Fig. 3. Schematic drawing showing the relation between spin, moment, and precession of a spinning body.

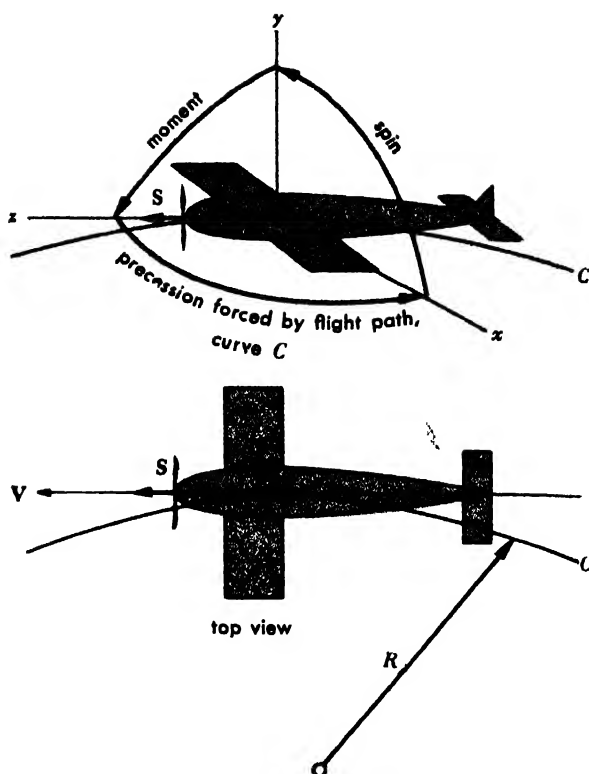


Fig. 4. Rotating elements such as a turbine rotor or propeller in an airplane following a curved flight path exert gyroscopic moments, as shown.

4). As the airplane follows the flight path curve  $C$ , it is forcing the rotating parts with spin  $S$  to precess with velocity  $\omega_p = V/R$ . From Eq. (3), then, the airplane must exert a moment  $M_z = S\omega_p I_z$ , where  $I_z$  is the moment of inertia about  $z$  of the rotating parts. If the elevators are not set to produce the required moment  $M_z$ , the nose of the airplane will rise.

All flight paths which involve sharp and rapid turns and loops will be accompanied by gyroscope moments exerted by the rotating parts, although such moments are usually small as compared with the aerodynamic moments acting. [R.E.BO.]

**Bibliography:** G. Joos, *Theoretical Physics*, 3d ed., 1958; A. G. Webster, *Dynamics of Particles and of Rigid, Elastic, and Fluid Bodies*, 2d ed., 1959.

## Precession of equinoxes

A slow change in the direction of orientation of Earth's axis of rotation which results in a gradual westward motion of the equinoxes. There are two types, known as lunisolar precession and planetary precession. The total precession is the sum of these two. The phenomenon of lunisolar precession, which is by far the more important, was discovered by Hipparchus about 125 B.C., and was first explained by Sir Isaac Newton.

The term equinox has a dual meaning. It refers to (1) either of the two imaginary points at which the ecliptic (the apparent annual path of the Sun among the stars) crosses the celestial equator (the

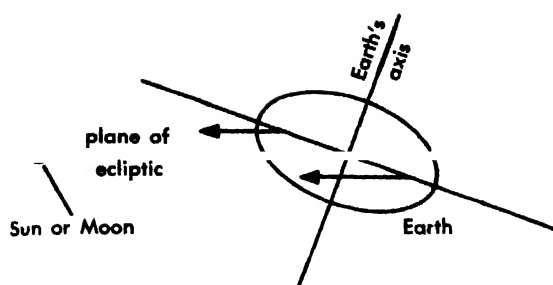


Fig. 1. The gravitational attraction of the Sun or Moon on Earth results in a torque which tends to tip Earth's axis.

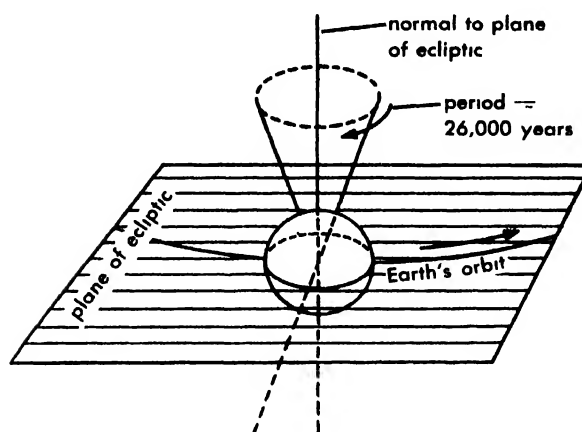


Fig. 2. Conical motion of Earth's axis. One complete cycle of precession requires 26,000 years.

great circle in which the plane of Earth's equator intersects the celestial sphere), or (2) the date when such a crossing occurs. During its annual journey around the ecliptic the Sun crosses the celestial equator twice. There are therefore two equinoxes each year; the vernal equinox occurs about March 21, and the autumnal equinox occurs about September 23. The annual westward motion of the equinoxes caused by precession amounts to about  $50''.27$ . The general motion of Earth's axis of rotation consists of precession plus a second angular motion of the axis which is called nutation. For general discussions of these phenomena, see NUTATION (ASTRONOMY AND MECHANICS); PRECESSION. See also ASTRONOMICAL COORDINATE SYSTEMS; EARTH, ORBITAL MOTION; EQUINOX.

**Physical causes.** The lunisolar precession of the equinoxes is caused by the gravitational attraction of the Sun and the Moon which, as a result of the polar flattening of Earth and the inclination of Earth's axis, gives rise to a small turning moment, or torque, on the Earth in its orbit (Fig. 1). This torque is not constant because of the varying positions in space of the three bodies involved. Because of the closer distance of the Moon, the magnitude of the torque due to the Moon is greater than that due to the Sun. As a result of this torque, Earth's axis describes a cone about the normal to the plane of its orbit (Fig. 2); the period of this

precession is approximately 26,000 years in a direction opposite to that of Earth's rotation, that is, the equinoxes precess from east to west. Because the angular momentum of the spinning Earth is so large, the rate of precession is extremely slow.

**Planetary precession.** This is a comparatively small eastward motion of the equinoxes caused by the action of the other planets in altering the plane of Earth's orbit. Planetary precession also causes a small change in the rate of lunisolar precession.

**Effects.** As a result of precession, Polaris will not always be the pole star. Vega will be nearer to the north celestial pole in about 12,000 years;  $\alpha$  Draconis was the pole star about 4600 years ago. Another effect of precession is that the signs of the zodiac no longer correspond to their respective constellations. The sign of Aries is now in the constellation of Pisces, and so on [K.W.P.]

**Bibliography:** H. N. Russell, R. S. Dugan, and J. Q. Stewart, *Astronomy*, vol. 1, 1945.

## Precious stones

The materials found in nature that are used frequently as gem stones, including amber, beryl (emerald and aquamarine), chrysoberyl (cat's-eye and alexandrite), coral, corundum (ruby and sapphire), diamond, feldspar (moonstone and amazonite), garnet (almandine, demantoid and pyrope), jade (jadeite and nephrite), jet, lapis lazuli, malachite, opal, pearl, peridot, quartz (amethyst, citrine, and agate), spinel, spodumene (kunzite), topaz, tourmaline, turquoise and zircon. See GEM.

The terms precious and semiprecious have been used to differentiate between gem stones on a basis of relative value. Because there is a continuous gradation of values from materials sold by the pound to those valued at many thousands of dollars per carat, and because the same mineral may furnish both, a division is essentially meaningless. All but a few of the gems listed cost many dollars per carat in fine quality. Gold, the symbol of concentrated value and the measure of preciousness, brings less than \$0.25 per carat. Thus, almost all gem stones merit the use of the term precious.

[R.T.L.]

## Precipitation (chemistry)

The process of producing a separable solid phase within a liquid medium. In a broad sense, precipitation represents the formation of a new condensed phase, although other terms are often used to describe the process. Thus, (1) a vapor or gas condenses to liquid droplets, or more specifically as in meteorology, water vapor in the atmosphere precipitates to form rain, snow, or ice; (2) a substance in the liquid state freezes or solidifies; (3) a dissolved component crystallizes from a supersaturated solution; (4) a new solid phase gradually precipitates within a solid alloy as the result of a slow, inner chemical reaction; or (5) a metal electrodeposits upon the passage of an electrical current through a solution.

In analytical chemistry, precipitation is widely used to effect the separation of a solid phase in an aqueous solution. For example, the addition of a water solution of silver nitrate to a water solution of sodium chloride results in the formation of insoluble silver chloride. Quite often, one of the components in the solution is thus virtually completely separated in a relatively pure form. It can then be isolated from the solution phase by filtration or centrifugation, and the substance determined by weighing. This procedure is known as gravimetric analysis. Precipitation may also be used merely to effect partial or complete separation of a substance for purposes other than that of gravimetric analysis. Such purposes might involve either the isolation of a relatively pure substance or the removal of undesirable components of the solution.

**Solubility-product constant.** The extent to which a component can be separated from solution can be determined from the solubility-product constant obtained by determining the quantity of dissolved substance present in a known amount of saturated solution. This value is known as the solubility. The solubility can be drastically altered merely by adding to the solution any of the ions comprising the precipitate, as for example, the addition of varying quantities of either silver nitrate or sodium chloride to a saturated solution of silver chloride. Although solubility can be altered over a wide range, the solubility product itself remains practically constant over this same range.

The solubility-product constant can be used to ascertain the quantity of dissolved component remaining unprecipitated in the presence of known concentrations of the ions common to the precipitate. By proper adjustment of the concentration of the added common ion, it is possible to reduce the quantity of dissolved component to a negligible value, although never to zero. It is an extremely important criterion of a method of gravimetric analysis that the quantity of unprecipitated component be negligible, particularly in comparison with the quantity of precipitate formed. The analytical chemist uses the word quantitative to describe such a chemical reaction. See SOLUBILITY PRODUCT CONSTANT.

**Impurities.** The purity of the precipitated solid phase is of major concern, both for the preparation of a desired chemical compound and for a quantitative method of gravimetric analysis. It is not possible for a precipitate to be formed as an absolutely pure compound by chemical reaction within the solution phase. Other soluble substances present in the solution, such as the ions not involved in the structure of the precipitate, tend to accompany the solid phase in varying amounts. This phenomenon is known as coprecipitation. The fraction of the total quantity of such foreign ions coprecipitating may be quite small. Although this fraction depends on experimental variables, it is highly dependent upon the relationship between the solubility characteristics of the desired chemical precipitate and

the foreign substance. As a specific example, partial precipitation of iodide (as silver iodide) using silver nitrate would coprecipitate (as silver chloride) only a small fraction of any chloride present, whereas it would coprecipitate (as silver bromide) a larger fraction of any bromide present. The iodide is more insoluble than the bromide, which is more insoluble than the chloride. Knowledge of the relative solubility characteristics of the chemical species present is thus extremely desirable to the chemist who needs to know whether foreign substances are being collected with the solid or are being left in solution.

A precipitated phase may incorporate foreign ions within its structure in several ways. Best understood of these is isomorphous mixed-crystal formation. Radium and barium sulfates form isomorphous mixed crystals because the two compounds have the same crystal structure and the ionic radii of radium and barium are not greatly different. Thus, radium and barium ions are interchangeable within the crystal lattice to a considerable degree. Silver bromide and silver chloride will also form isomorphous mixed crystals. Because of the ease with which interchange can take place within the crystal lattice, isomorphous mixed crystal systems should be avoided if a good separation of the two different ionic species is desired. On the other hand, the property of isomorphism may also be put to good use in concentrating minute traces. For example, barium sulfate precipitated in the presence of minute traces of radium will carry with it almost all of the radium. Such procedures are frequently used in the collection of minute traces of radioactive species.

An ion present at high dilution may be incorporated, apparently by mixed crystal formation, even though such formation would not be predicted on the basis of crystallography and ionic radii. An example of this is the coprecipitation of traces of lead with potassium chloride. This phenomenon is known as anomalous mixed crystal formation, or isodimorphism.

When foreign-ion incorporation cannot be ascribed to isomorphism, coprecipitation may occur by adsorption. Residual charges at the surface of a precipitate attract charged ions in the solution. The adsorption process results in a greater concentration of foreign ions near the precipitate surface than exists in the main body of the solution. Adsorbed foreign ions may remain quite firmly attached to the solid. In fact, they may be covered as succeeding layers of the crystal are deposited and cause imperfections within it. This latter phenomenon is known as occlusion, although it arises as a result of adsorption. The term occlusion should be distinguished from inclusion, which refers to the mechanical trapping of the solutions (and solutes in it) within the precipitate.

Sometimes one substance in a mixture precipitates rapidly but a second foreign substance then precipitates slowly as a second solid phase. This is not generally considered to be coprecipitation.

but is referred to as postprecipitation. The postprecipitation of zinc sulfide with copper sulfide is a typical example.

**Methods for reducing contamination.** In an effort to reduce contamination by foreign ions, the chemist resorts to various techniques. Precipitation from dilute solution is often effective. Heating the reaction mixture, that is, digesting the solid in the liquid phase, speeds recrystallization processes by which incorporated foreign ions may be returned to the solution phase. Precipitation from homogeneous solution, a technique in which the desired precipitating reagent is formed internally within the solution by chemical synthesis, results in the slow formation of large crystals of small surface area, and hence, lessens coprecipitation.

If all these methods fail to reduce adequately the quantity of foreign ions incorporated in the solid phase, then reprecipitation is applied. The precipitate is dissolved and reprecipitated by the previous procedure. As in the initial precipitation, a major fraction of the foreign ions will remain in solution. The process of reprecipitation must be repeated until the quantity of foreign ions present in the precipitate can be disregarded. See ADSORPTION; CRYSTALLIZATION; ELECTRODEPOSITION ANALYSIS; GRAVIMETRIC ANALYSIS; ISOMORPHISM (CRYSTALLOGRAPHY); NUCLEATION; SATURATION OF SOLUTIONS; SEPARATION (CHEMICAL AND PHYSICAL).

#### PRECIPITATION FROM HOMOGENEOUS SOLUTION

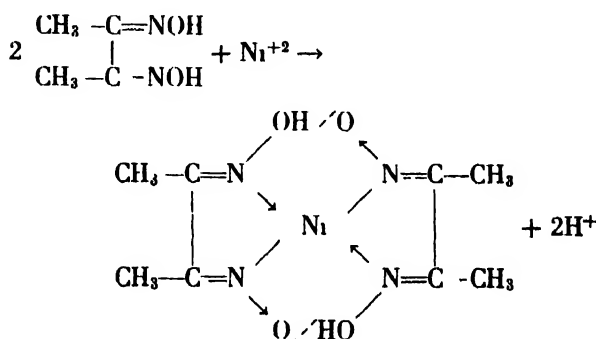
In this technique, the precipitant is generated *in situ* within the solution phase instead of being added directly, as in the conventional manner.

The substances required to effect precipitation from homogeneous solution can be generated within a solution phase in a variety of ways. For example, if a source of hydroxyl ions is needed to precipitate metallic ions, urea in solution can be hydrolyzed to produce ammonium hydroxide to act as the source. Other anions needed as precipitants can often be formed by the hydrolysis of appropri-

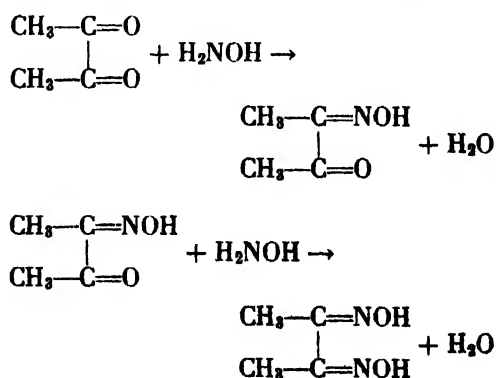
ate esters. Thus, dimethyloxalate can serve as a source of oxalate ions to precipitate thorium, and dimethyl sulfate can be the source of sulfate ions to precipitate lead. There also are methods for the generation of cations. For example, silver ions can be slowly released from a complex ion such as  $\text{Ag}(\text{NH}_3)_2^+$  to precipitate chloride ions. Various investigators have developed ingenious methods for producing the necessary precipitant to effect precipitation from homogeneous solution.

The formation of larger, more perfect, and purer crystals occurs almost without exception when precipitation from homogeneous solution is used in place of direct addition of the precipitant.

The photomicrographs show the difference in appearance of nickel dimethylglyoximate precipitated by two different methods. The precipitate shown in part *b* of the illustration was formed by the direct addition of dimethylglyoxime to a solution containing nickel ions, as follows:



The more perfect crystals shown in part *a* were precipitated from homogeneous solution by synthesizing the necessary dimethylglyoxime from biacetyl and hydroxylamine by the following reactions:



Because the rate at which a precipitant is generated can be controlled very closely, precipitation from homogeneous solution is used as a research technique in studies of the mechanisms of nucleation, precipitation, and coprecipitation. The technique has many applications in gravimetric analysis, in the production of pure chemicals, and in the control of particle size of crystals. [L.C.]

**Bibliography:** L. Gordon, M. L. Salutsky, and H. H. Willard, *Precipitation from Homogeneous*



Nickel dimethylglyoximate precipitated by two different methods. (a) Precipitation from homogeneous solution. (b) Direct addition. (From L. Gordon and E. D. Salesin, *Precipitation from homogeneous solution*, *J. Chem. Educ.*, 38(1):16, 1961)

*Solution*, 1959; L. Gordon and E. D. Salesin, Precipitation from homogeneous solution, *J. Chem. Educ.*, 38(1):16, 1961.

## Precipitation (meteorology)

The fallout of water drops or frozen particles from the atmosphere. Liquid types are rain or drizzle, and frozen types are snow, hail, small hail ice pellets (also called ice grains; in the United States, sleet), snow pellets (graupel, soft hail), snow grains, ice needles, and ice crystals. In England sleet is defined as a mixture of rain and snow, or melting snow. Deposits of dew, frost, or rime, and moisture collected from fog are occasionally also classed as precipitation.

All precipitation types are called hydrometeors, of which additional forms are clouds, fog, wet haze, mist, blowing snow, and spray. Whenever rain or drizzle freezes on contact with the ground to form a solid coating of ice it is called freezing rain, freezing drizzle, or glazed frost; it is also called an ice storm or a glaze storm, and sometimes is popularly known as silver thaw or erroneously as a sleet storm.

Most precipitation particles carry an electrostatic charge, either positive or negative, but the origin of the charge and its relation to other problems of atmospheric electricity are not completely understood. See ATMOSPHERIC ELECTRICITY; THUNDERSTORM.

Rain, snow, or ice pellets may fall steadily or in showers. Steady precipitation may be intermittent though lacking sudden bursts of intensity. Hail, small hail, and snow pellets occur only in showers; drizzle, snow grains, and ice crystals occur as steady precipitation. Showers originate from instability clouds of the cumulus family, whereas

steady precipitation comes from stratiform clouds. See CLOUD.

Depth of rain, melted snow, or other forms of precipitation is measured at the ground and is referred to as precipitation, precipitation amount, or simply as rainfall. Frozen forms are first melted to obtain their water content, and then the amounts are recorded in millimeters (mm) or inches and hundredths of liquid water. Separate measurements are made of the depth of unmelted snow, hail, or other frozen forms. See PRECIPITATION GAGES.

**Liquid types.** Rain and drizzle are somewhat arbitrarily differentiated. Raindrops are generally over 0.5 mm in diameter and differ from drizzle mainly in having larger sizes. Drizzle may also be distinguished from rain by the meteorological conditions of its formation; drizzle falls usually from fog or thick stratus clouds, whereas rain comes from clouds of cumulus type or clouds extending well above the freezing level. Raindrops are rarely larger than 5 mm in diameter because larger sizes tend to break up.

The small raindrops are approximately spherical, but large falling drops are flattened (see Fig. 1) especially on the bottom side. The following speeds of falling drops, in meters per second, were measured in still air by R. Gunn (drop diameters are in millimeters):

Speed	0.27	2.06	4.03	6.49	8.06	8.83
Diameter	0.1	0.5	1.0	2.0	3.0	4.0

**Frozen precipitation.** Snowflakes are branched, six-point star crystals, irregular forms such as matted ice needles, or combinations of both; often they are coated by rime. The speed of falling flakes is variable; it is greater when the flakes are rimed and is mostly 1–2 meters/sec. See SNOW.

Hail, mostly seen in thunderstorms, forms with the aid of warm, moist air when clouds build to great heights. Hailstones usually have concentric layers of rime and hard ice, typically 1 cm in diameter but ranging up to more than 10 cm. See HAIL.

Ice pellets are of hard ice, either clear or opaque, and may be spherical or irregular. They are formed by the freezing of rain or drizzle drops.

Snow pellets are of soft, opaque ice, irregularly shaped, sometimes with scalloped edges and often studded with a few oblong or branched crystals. Typical diameters are 2–5 mm. They fall usually in showers, sometimes even when ground temperatures are a few degrees above freezing. U. Nakaya found their specific gravity to be about 0.12. Snow grains are similar but smaller and flatter, with diameters around 1 mm or less; they fall generally in small amounts from stratus clouds or fog.

Ice needles are narrow, pointed crystals, roughly 1–3 mm long and about  $\frac{1}{4}$  mm in diameter, which fall singly or in clusters, mostly at temperatures near or a little below freezing.

Ice crystals are small rods or plates. They fall in cold, stable air that is often without clouds, and glisten in light as they settle to the ground.

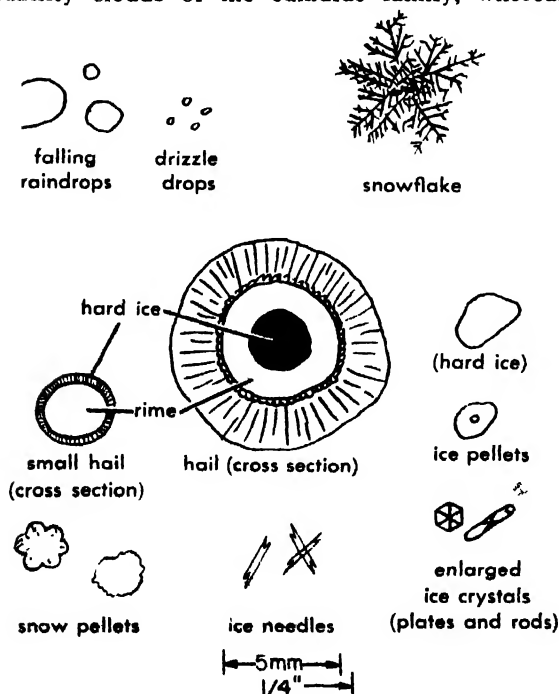


Fig. 1. Various forms of precipitation.

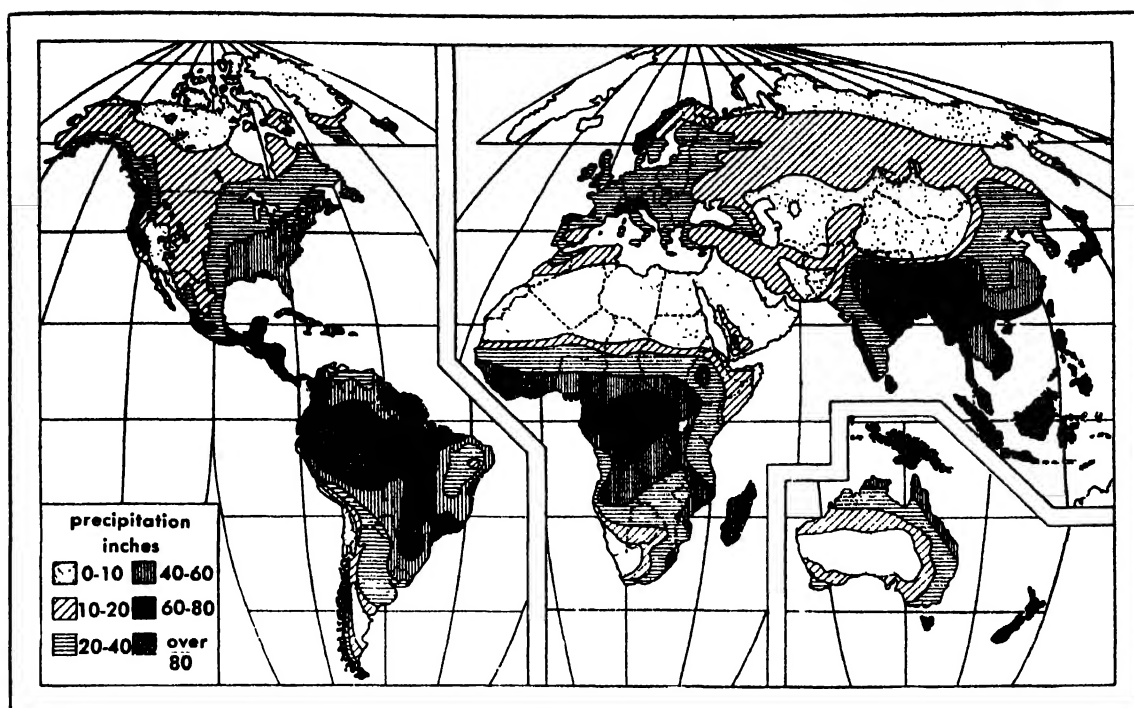


Fig. 2. Distribution of average annual precipitation over the land areas. (From N. A. Bengtson and W. Van

Royen, *Fundamentals of Economic Geography*, 4th ed., Prentice-Hall, 1956)

**Geographical distribution.** Precipitation is part of the hydrologic cycle, the continuous interchange of water between sea, land, and atmosphere, but its distribution over the earth is uneven (Fig. 2).

Some world-record rainfall amounts are, for 1 min, 1.23 in., Unionville, Md., 1956; 1 hour, 12.00 in., Holt, Mo., 1947; 24 hours, 45.99 in., Baguio, Philippines, 1911; 2 days, 65.79 in., Funkiko, Formosa, 1913; 31 days, 366.14 in., Cherrapunji, India, 1861; 1 year, 1,041.78 in., and 2 years, 1,605.05 in., both Cherrapunji, India, 1860-1861 (data from U.S. Weather Bureau).

**Precipitation and weather.** Precipitation occurs most often in cyclones or tropical disturbances. In weather forecasting, several synoptic (weather map) types of precipitation are recognized: warm front, warm moist air rising over a wedge of cold air; cold front, cold air undercutting and lifting warmer air; convective, caused by local updrafts of moist air; convergent, general lifting of air caused by convergence at low levels; and orographic, moist air forced upward on mountain slopes. Any type may act alone; however, convection or convergence may occur with other types. Lake snow falls when air, at first extremely cold, blows off a bay or off lakes such as the Great Lakes. Minor causes of precipitation are turbulence, contact cooling (air moving over a colder surface), and nighttime radiation from cloud tops, but the amounts from these sources are small.

**Condensation in the atmosphere.** Basically this requires that air be cooled to and below its dew point. Then, minute water droplets form by condensation, or ice crystals form by sublimation. These droplets or ice crystals are cloud particles

from which, by a growth mechanism, the larger precipitation particles may form. The only cooling process sufficient to produce appreciable precipitation is adiabatic expansion, which occurs in air rising toward lower pressures (see **ATMOSPHERIC ADIABATIC CHANGE**). The rate of condensation in rising saturated air is greater the warmer the air and is directly proportional to the speed of ascent. For discussions of condensation nuclei, formation of precipitation in clouds, and artificial stimulation of precipitation, see **CLOUD PHYSICS**; **WEATHER MODIFICATION**.

For discussions of other topics related to precipitation see **DEW**; **DEW POINT**; **FOG**; **HUMIDITY**; **HYDROLOGY**; **RAIN SHADOW**; **VAPOR PRESSURE**.

[J.R.F.]

**Bibliography:** T. F. Malone (ed.), *Compendium of Meteorology*, 1951; S. Petterssen, *Weather Analysis and Forecasting*: vol. 2, 2d ed., 1956.

## Precipitation gages

Instruments used to measure the amount of rainfall or snowfall expressed as inches or centimeters depth of water which falls on a level surface. When precipitation gages are equipped with a recorder, the time of occurrence and the rate or intensity are available as well. Other forms of precipitation, such as dew, frost, and moisture absorbed by the soil, are also of considerable interest, but they are not measured by the precipitation gages routinely used by the meteorological services.

**General design features.** The gage is basically an open container and funnel constructed to minimize any splashing out and mounted 1-3 ft above the surface to prevent splashing into the container,



yet not so high that its catch is affected by the wind which normally increases in velocity with height. From the volume of water calibrated and the area of the opening, the depth of water is easily obtained. Various schemes are used to obtain the water volume. The U.S. Weather Bureau calibrates the catch in a cylinder whose area is one-tenth that of the gage area. The depth is measured by a dip stick calibrated in inches. Gages used in other countries have bottles, graduated cylinders, or other arrangements for obtaining the volume of the catch. The funnel in each of these gages reduces water loss by evaporation.

**Recording instruments.** A continuous type of recording rain gage uses a spring balance to record on a clock-driven chart the weight of the precipitation calibrated. High sensitivity is obtained by using a slotted linkage so that large amounts of rainfall cause the recorder pen to travel back and forth several times across the chart. Another type of recording gage uses a shallow V-shaped two-compartment tipping bucket, pivoted at the vertex

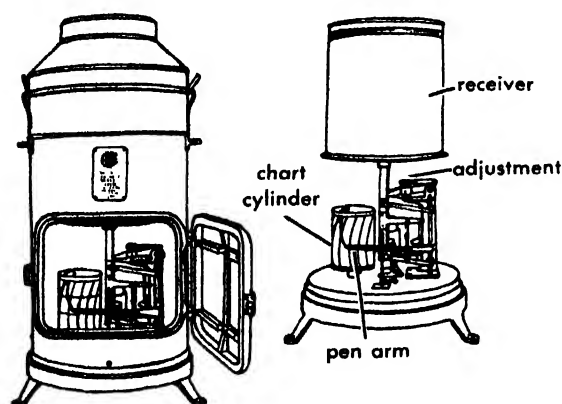


Fig. 1. Fergusson weighing and recording type of gage. (From F. A. Berry, Jr., E. Ballay, and N. R. Beers, eds., *Handbook of Meteorology*, McGraw-Hill, 1945)

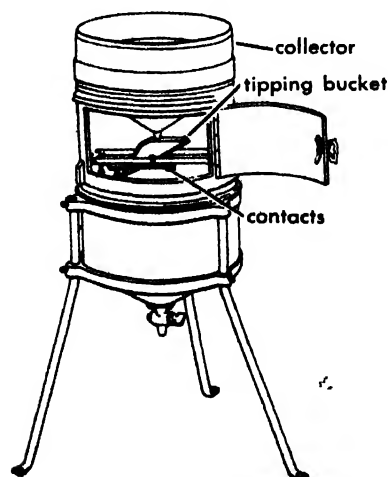


Fig. 2. A tipping-bucket gage. (From F. A. Berry, Jr., E. Ballay, and N. R. Beers, eds., *Handbook of Meteorology*, McGraw-Hill, 1945)

so that it has two stable positions. The vertical partition lies just below the opening in the funnel and directs the flow into the other position where the second compartment fills while the first empties into a container. Each tip of the bucket corresponds to 0.01 in. of rainfall, and closes an electrical contact to actuate a remote register. Still other types of gages use a float to actuate the recording pen. Some of these have an automatic siphon to empty the float chamber when a specific amount of rainfall (0.02 in.) is collected.

**Gaging snowfall.** When a suitable wind screen is used, snowfall can be measured by melting the snow collected in a rain gage (the funnel must be removed). Fallen snow is also measured by taking sample cores. A device for measuring the mass of snowfall uses a cobalt-60  $\gamma$ -ray source at ground level columnated toward a Geiger counter mounted several feet above the surface. As snow covers the surface, the reduction in counting rate as a result of  $\gamma$ -ray absorption by the snow is used to determine the mass of snow in the column. An attractive feature of the device is that the counts are easily telemetered by radio, making it very useful for mountainous regions. See SNOW GAGE; SNOW SURVEYING. [V. E. SUOMI]

**Bibliography:** F. A. Berry, Jr. et al. (eds.), *Handbook of Meteorology*, 1945; W. E. K. Middleton, *Meteorological Instruments*, 3d ed., 1953.

## Precipitin

A term used in serology to describe an antibody that reacts with its corresponding antigen to give a visible precipitate. Since precipitin reactions are given by some purified antibody solutions that also give agglutination and other serological reactions, this is purely an operational definition. Incomplete antibodies are also known that do not precipitate with soluble antigens although they will agglutinate and otherwise add on to particulate antigens, and evidence for their occurrence together with precipitating antibody can be obtained for most sera. Precipitins may be quantitated by noting the endpoint dilution (titer) of serum required to give a precipitate at the threshold of visibility, or the amount of antibody may be determined in milligrams or micrograms by appropriate analysis of the precipitate, with correction for the antigen contained therein. See ANTIBODY; ANTIGEN; PRECIPITIN TEST. [H. P. TREFFERS]

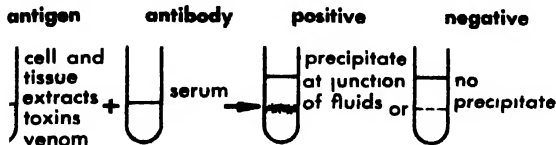
**Bibliography:** S. Raffel, *Immunity*, 2d ed., 1961; G. S. Wilson and A. A. Miles (eds.), *Topley and Wilson's Principles of Bacteriology and Immunity*, 2 vols., 5th ed., 1964.

## Precipitin test

A term used in serology to describe a specific reaction between antigen and antibody that results in a visible precipitate.

The precipitin test or reaction is a technique used in serology to determine either the antigen or its corresponding antibody. This reaction differs

from agglutination chiefly in the size of particles involved. In agglutination, the particles are usually cells and are usually large enough to be seen under the microscope. The precipitin test is most often used as a test for particular antigens in extracts of unknown bacteria, blood stains, or tissues, and it finds application in the diagnosis of bacterial diseases, such as streptococcal infections and anthrax, and in the medical-legal identification of the animal species providing a blood stain or tissue. The most important identification of antibody is that result-



The ring test, one form of the precipitin test (From C. J. Witton, *Microbiology with Applications to Nursing*, 2d ed., McGraw-Hill, 1956)

ing from syphilitic infection, although the antigen utilized may be derived from normal as well as syphilitic tissue. The Kahn, Kline, and other flocculation tests are examples. The precipitin test may be initiated by mixing clear antigen and antibody solutions and observing whether a precipitate forms after a short period at 37° or after hours or days at 0°C. It may be quantitated by observing the limiting dilution of antibody, or less desirably, the antigen, at which the reaction becomes just visible. Or the amount of precipitate may be determined as milligrams or micrograms of protein by chemical or instrumental procedures. These may estimate as little as 0.1  $\mu$ g antibody nitrogen. In other variations, the antigen solution is laid over serum containing the antibody. A visible precipitate layer at the interface is a positive test. Antigens and antibodies may also be permitted to diffuse toward one another through agar gels. Antigen solutions diffusing into the agar form a distinct and stable band for each component which has antibody present. With this technique, purified and supposedly homogeneous antigens have been shown to consist of seven distinct components. See AGGLUTINATION REACTION; ANTHRAX; ANTIBODY; ANTIGEN; SYPHILIS.

[H. P. TREFFERS]

**Bibliography:** J. F. Ackroyd (ed.), *Immunological Methods*, 1964; E. A. Kabat and M. M. Mayer, *Experimental Immunochemistry*, 2d ed., 1961.

## Precision approach radar (PAR)

A radar system located on an airfield for observation of the position of an aircraft with respect to an approach path and specifically intended to provide guidance to the aircraft during its approach to the field.

The PAR system consists of a ground radar equipment which is alternately connected to two antenna systems. One of these antenna systems

sweeps a narrow beam over a 20° sector in the horizontal plane. The second antenna system sweeps a narrow beam over a 7° sector in the vertical plane. Since the reflected energy from the aircraft produces information on the location of the aircraft in terms of angles and distance, the two sweeps give the three-dimensional location of the aircraft (see Fig. 1).

To obtain the necessary discrimination, the antenna systems must produce extremely narrow beam widths. Further, since it is necessary to obtain information very often, these antennas must be capable of scanning at a very fast rate (about 4 scans/sec). The desired performance is obtained by use of long linear arrays of dipoles. Operating on a frequency of approximately 10,000 megacycles, some of these antenna arrays have employed 130 dipoles and had lengths of approximately 85 ft. The resulting beam width for the antenna sweeping in the horizontal plane is a height of about 1.5° and a width of 0.6°. The antenna sweeping in the vertical plane has a width of approximately 30° and a height of 0.4° (see Fig. 2). Rapid sweeping is accomplished by changing the phasing of the energy fed to the dipoles. This phase change is accomplished by varying the dimensions of the wave guide that feeds the dipoles.

Various presentations of the data from the PAR have been produced. One of the simplest presentations employs two cathode-ray oscilloscopes. One of these presents azimuth versus distance, the other elevation versus distance (see Fig. 3). The operator positions transparent plastic scales over the spots representing the target. Attached to the scales are cam mechanisms that rotate voltage dividers. The cams are cut with curves that correspond to the localizer and glide-slope positions that an aircraft should follow in making an optimum approach. With an aircraft making an optimum approach, the cams always position the arm on the voltage divider so that the output is zero.



Fig. 1. Guidance with precision approach radar. (From P. C. Sandretto, *Electronic Aviation Engineering*, International Telephone and Telegraph Corporation, 1958)

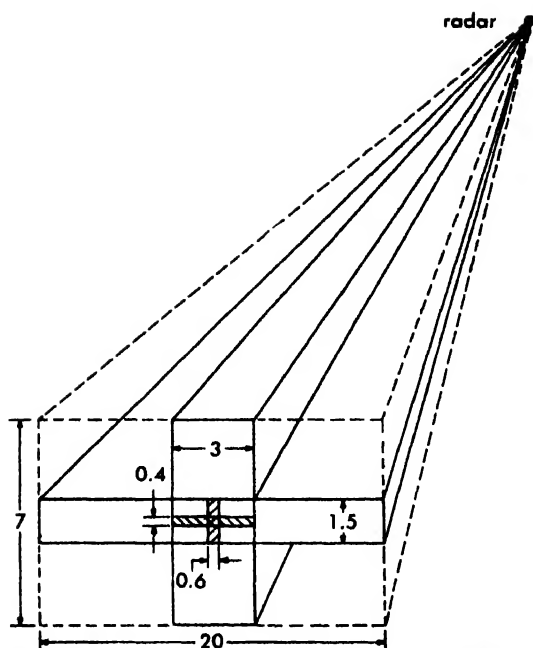


Fig. 2. Field patterns of the antennas of the radar low-approach system. The dimensions shown are all in degrees. The broken lines indicate the total movement of the field patterns. The cross-hatched areas represent the dimension of the field patterns from the antennas when they are not sweeping.

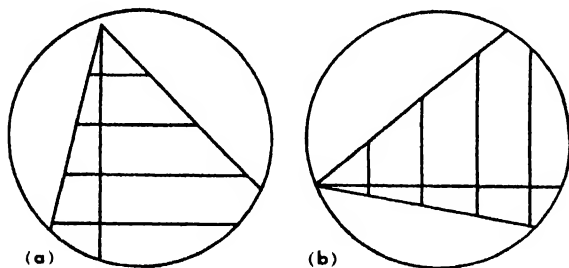


Fig. 3. Simple display of a low-approach radar. (a) Azimuth-distance. (b) Elevation-distance.

Deviation above or below or to the right or the left of the optimum path produces a positive or negative output voltage. This voltage is connected to a cross-pointer instrument such as that used with ILS. See INSTRUMENT LANDING SYSTEM (ILS).

Probably the most elaborate PAR presentation is that known as the AZ-EL scope. On this oscilloscope presentation, all information is generated electronically. The upper half of the oscilloscope is devoted to the presentation of glide-slope (vertical position) indication and the lower half of the scope is devoted to azimuth (localizer) indication. The optimum paths in the horizontal and vertical planes are indicated by electronically generated lines. Distance is displayed logarithmically so that about half of the scale shows a distance equal to 3 miles while the remaining half is equal to about 7 miles.

It is not necessary to have specific airborne equipment in order to utilize the PAR system, although such equipment has been used experimentally. In common practice, the ground PAR operators call instructions to the pilot, indicating the relative position of the aircraft with respect to the optimum approach path or give the pilot maneuvering instructions that will bring the aircraft to the approach path. See AIRCRAFT LOW-APPROACH SYSTEMS; NAVIGATION SYSTEMS, ELECTRONIC. [P.C.S.]

*Bibliography:* P. C. Sandretto, *Electronic Aviation Engineering*, 1958.

## Pregnancy

In humans, the period of about 280 days required for the normal intrauterine development of a child. Fertilization, or conception, occurs if, shortly after ovulation, a male sperm penetrates the ovum. Ovulation is usually midway between menstrual periods. The ovum passes into the Fallopian tube and then into the uterus. Most fertilization is believed to occur in the tube within a day or so of ovulation. If the fertilized ovum does not proceed into the uterus, one form of ectopic, or displaced, pregnancy may occur. See FERTILIZATION; OVUM; SPERM CELL.

Under normal conditions, the fertilized ovum will pass into the uterine cavity and become implanted in the lining, or mucosa, called the endometrium. This highly vascular tissue, which usually responds to the cyclic regulation of female hormones, now gradually comes under the influence of the gestational or placental hormones. These tend to maintain the mucosa, instead of allowing it to be sloughed off periodically. In addition, they directly and indirectly exert influences on the entire endocrine system so that almost every tissue and organ of the body is affected. See PLACENTATION.

As the embryo develops, rapid differentiation of germ layers into organs and systems occurs, accompanied by a corresponding growth in the size of the embryo and of the supporting placenta.

**Maternal changes.** Changes in the mother are most marked in the reproductive organs, including the uterus itself. A slight enlargement and softening of the uterus may be detectable on examination by the end of the second month. As further enlargement occurs, the muscular wall becomes hypertrophied and then stretched, so that it is soft and somewhat flexible. During the fourth month the uterus ordinarily rises above the pelvis and becomes abdominal. By the end of pregnancy the upper portion of the uterus lies just beneath the anterior rib margins.

Changes in the breasts begin early in pregnancy with enlargement and tenderness being common. As term approaches, the nipples and areolae increase in size and deepen in pigmentation, reflecting the proliferation of the glandular elements of the breasts. A watery or slightly turbid discharge, called colostrum, may be expressed from

the nipples, especially in women who have had previous pregnancies.

Changes in the vagina, ovaries, and cervix accompany the progression of pregnancy. Systemic alterations in renal function, cardiovascular adaptations, blood volume and cell counts, as well as in chemical constituents of blood and tissue all reflect the complex physiologic changes required for successful pregnancy.

**Ectopic and multiple pregnancy.** Ectopic pregnancy occurs when the fertilized ovum becomes implanted on tissues other than the endometrium of the uterus. This is said to happen in about 1 out of every 300 pregnancies and most of these aberrant implantations are found in the Fallopian tubes. Other sites of implantation are the ovaries and the abdominal cavity.

Multiple pregnancy indicates the simultaneous development of two or more fetuses. Twins occur about once in 90 pregnancies, triplets once in 10,000, and greater numbers are unusual enough to be widely heralded events.

**Termination of pregnancy.** Pregnancies may proceed to term or may be concluded by natural or artificial means prior to normal delivery. The term abortion is used to indicate the passage of a fetus of less than 1000 grams, or before the twenty-eighth week of pregnancy. This may occur spontaneously as a result of any one of many causes, including criminal induction.

When pregnancy is terminated after the twenty-eighth week and before the end of normal pregnancy, a premature infant is born. Premature labor commonly follows stress, infections, multiple pregnancy, premature rupture of the fetal membranes, and other specific and nonspecific events. About 5-7% of all pregnancies end in premature labor. The infant mortality decreases with the length of pregnancy and the size of the infant.

Although the normal period of pregnancy in humans is given as 280 days from the beginning of the last menstrual period, or 269 days from the apparent time of fertilization, no completely accurate method exists for this measurement. There are cases in which pregnancy has continued for 300 days or longer with no grossly abnormal features in the offspring. [E.C.ST.]

## Pregnancy, disorders of

The relatively common minor complaints of nausea, vomiting, fatigue, and constipation and also the more serious disorders mentioned here.

Pernicious vomiting is usually found early in pregnancy. Its cause is unknown but appropriate treatment is usually successful.

The toxemias of pregnancy are a group of diseases found in the last trimester or following delivery. They are characterized by hypertension, albuminuria, and edema. The convulsions and coma of eclampsia may ensue if treatment is delayed.

Chronic hypertension, heart disease, and chronic kidney disease are preexisting conditions which

produce a high percentage of complications during and after pregnancy.

Among the infectious diseases, tuberculosis and syphilis present hazards to both mother and the unborn infant. Although formerly a dreaded killer, childbirth fever (puerperal sepsis), a streptococcal infection, has largely been eliminated in modern facilities, but still exists in backward areas. See STREPTOCOCCUS; SYPHILIS; TUBERCULOSIS.

Hemolytic diseases, capable of producing severe damage to the red blood cells of either the mother or child, may be largely avoided, or at least anticipated, by proper blood testing. Particular attention is given to the Rh factor and similar genetic blood groups. See BLOOD GROUPS.

Hemorrhage in pregnancy may be due to many causes but by far the most common is abortion; less common causes are ectopic pregnancy, premature labor, and premature separation of the placenta from the uterus.

In ectopic pregnancy, a not infrequent occurrence, a fertilized ovum becomes implanted in a tissue other than the uterine mucosa. Most often the site of implantation is some part of the Fallopian tube which becomes stretched and inflamed as growth proceeds; rupture is not uncommon. In other cases, the fertilized ovum may be implanted on some part of the peritoneum or mesentery and produce some symptoms if growth proceeds. There are occasional reports of the delivery of a full-term, normal infant by abdominal operation from such a site. In most cases, however, it is believed that the embryo does not survive and is gradually absorbed by the body.

It is estimated that of every 100 fertilizations, only 70% will develop into term infants. Of the remainder, one-third are abnormal so that implantation cannot occur, one-third result in abnormal embryos which are aborted, and one-third are lost by spontaneous abortion from other causes. See FERTILIZATION. [E.C.ST.]

## Prehnite

A mineral sorosilicate, composition  $\text{Ca}_2\text{Al}_2\text{Si}_3\text{O}_{10}(\text{OH})_2$ , crystallizing in the orthorhombic system. Distinct crystals are rare and it occurs usually in reniform and stalactitic aggregates with crystalline surface. Hardness is 6-6½ on Mohs scale; specific gravity is 2.8-2.9. It has a vitreous luster with a light green to white color. Prehnite is characteristically found lining the cavities in basaltic rocks. It is associated with datolite, zeolites, calcite, and pectolite. Noted localities in the United States are at Patterson and Bergen Hill, New Jersey; Westfield, Massachusetts; and Farmington, Connecticut. See SILICATE MINERALS. [C.S.HU.]

## Press fit

A press fit has negative allowance; that is, the bore in the fitted member is smaller than the shaft which is pressed into the bore. Tight fits have slight negative allowance so that light pressure is re-

quired to assemble the parts; they are used for gears, pulleys, cranks, and rocker arms. Medium force fits have somewhat greater negative allowance and require considerable pressure for assembly; they are used for fastening locomotive wheels, car wheels, and motor armatures. See ALLOWANCE; FORCE FIT; SHRINK FIT. [P.H.B.]

## Pressure

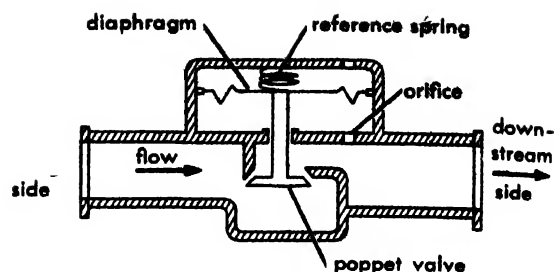
The ratio of force to area. The force per unit area at the interior of the Sun is estimated to be  $3 \times 10^{17}$  dyne/cm<sup>2</sup>. In interstellar space, pressure approaches zero. Atmospheric pressure at the surface of Earth is in the vicinity of 14 lb/in<sup>2</sup>. Pressures in enclosed containers less than this value are spoken of as vacuum pressures; for example, the vacuum pressure inside a cathode-ray tube is  $10^{-8}$  mm of Hg, meaning that the pressure is equal to the pressure that would be produced by a column of mercury, with no force acting above it, that is  $10^{-8}$  millimeters high. This is absolute pressure measured above zero pressure as a reference level. Inside a steam boiler, the pressure may be 800 lb/in.<sup>2</sup> or higher. Such pressure, measured above atmospheric pressure as a reference level is gage pressure, designated psig. See MILLIBAR. [F.H.R.]

## Pressure control, automatic

A form of feedback control system in which the controlled variable is the magnitude of the pressure of a fluid. In many chemical and petrochemical processes, maintenance of a pressure or vacuum is an important condition to successful operation. Pressure control is used not only to obtain a satisfactory reaction but also to obtain constant flow through a control valve or constriction when the main pressure source is variable. The configuration of a pressure control system is closed-loop; that is, the controlled pressure is measured and compared to some reference or set point. The difference between the set point and the actual pressure reading causes an actuator to open a valve. The valve admits more or less fluid, thereby increasing or decreasing the pressure at the point of measurement until a balance is reached. Some of the more complex systems adjust pressure by varying the speed of a pump or by decreasing pumping efficiency, but most use throttling methods. The same general principles which apply to the design and performance of all dynamical feedback control systems apply to pressure control systems.

These automatic pressure control systems require transducers that measure the pressure and convert the reading into some readily usable form, such as force, torque, rotation, displacement, electric voltage or current. Once converted into a usable form, the signal is processed in a controller and applied to an actuator which varies the pressure adjusting mechanism.

Pressure control systems have many possible forms. A common mechanical form is the reducing



Pressure regulating valve.

or regulating valve, which combines the functions of the pressure transducer, controller, and actuator in the simple device shown in the figure. The controlled pressure on the downstream side is adjusted by opening or closing a poppet valve which throttles the flow of fluid. As the valve is adjusted, the controlled pressure falls or rises. The poppet valve may be regarded as the actuator in a feedback control system.

The reference input is the force created by the compression of the spring. This force is compared to the force created by the pressure on the diaphragm and, if an unbalance exists, will open or shut the valve depending on whether the pressure force is lower or higher than the spring force. As in all feedback control systems, the possibility of self-induced oscillation exists. This would occur if change in valve area were too large for the error in pressure. Oscillation is prevented by restricting the feedback passage. An orifice, in combination with the volume of the diaphragm cavity, acts as a low-pass, low-gain element. This element performs the functions of a controller in preventing oscillation and producing satisfactory performance in the presence of disturbances.

The automatic pressure control system illustrated in this example is a simple one in which the various functions of pressure sensing, error computation, controlling, and actuation are carried out in a compact device. Simpler devices such as pressure relief valves which exhaust fluid into the atmosphere or overflow sump, or more complex systems using pressure transducers, conventional pneumatic or electric controllers, and separate actuators are also used, depending on the application.

The basic theory which describes the performance and design of pressure control systems is identical to that of any dynamical feedback control system. Considerations of stability, steady-state accuracy, sensitivity to extraneous variations, such as temperature and supply pressure, and response to random fluctuations are important. See CONTROL SYSTEM; PROCESS CONTROL. [J.R.R.]

## Pressure measurement

The determination of the magnitude of a (fluid) force applied to a unit area.

Pressure measurements are generally classified as gage pressure or absolute pressure. Gage pres-

sure is the difference between a given pressure and the pressure of the atmosphere. Absolute pressure is the total pressure, including that of the atmosphere. Atmospheric pressure was the first pressure that was really measured; it remains an important third category. Pressures less than atmospheric pressure are called vacuum (see VACUUM MEASUREMENT).

The table compares eight common units of pressure measurement. To avoid confusion, absolute or gage is often suffixed, for example, 100 pounds per square inch gage pressure (100 psig) or 115 pounds per square inch absolute pressure (115 psia).

In the laboratory, pressure is an important measurement, since the pressure level has a significant effect on most physical, chemical, and biological processes.

In industry—particularly in the process industries—pressure is measured and controlled to maintain uniformity of product, to guide in safe operation, to determine pumping head for fluid transfer, and to measure other variables indirectly, including weight, liquid level, temperature, flow and density of fluids, and hydraulic forces.

Pressure gages generally fall in one of four categories, based on the principle of operation: liquid columns, bell instruments, expansible-element gages, and electrical pressure transducers.

**Liquid-column gage.** This type of pressure gage includes barometers and manometers. It consists of a U-shaped tube partly filled with a nonvolatile liquid. Water and mercury are the two most common liquids used in this type of gage. See BAROMETER; MANOMETER.

If one leg is left open to the atmosphere, the difference in level is a direct measure of gage pressure. If two pressures are to be compared, one is applied to each leg of the gage. The level rises in the low-pressure leg and drops in the high-pressure leg. The differential pressure is the difference in level (head) multiplied by the density of the instrument liquid.

**Bell-type gage.** This consists of a housing in which a bell, open downward, is partially immersed in a liquid such as light oil or mercury. The liquid

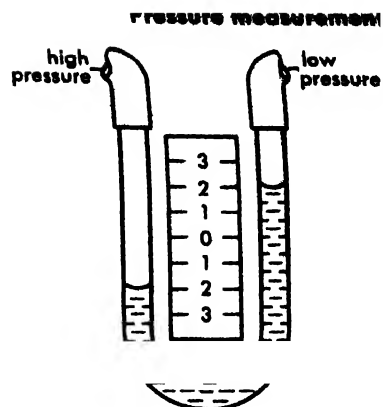


Fig. 1. Liquid-column gage (U-tube manometer).

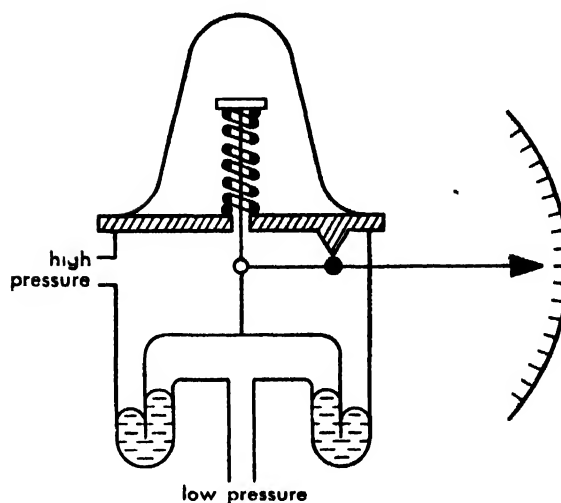


Fig. 2. Bell-type gage.

acts as a seal between pressure applied on one side of the bell and the atmosphere on the other side. Vertical motion of the bell, opposed by a calibrated coiled-wire spring, is a direct measure of gage pressure.

The bell-type gage can also be employed to measure the difference between two unknown pressures. Both sides of the bell are enclosed, and motion of the bell is detected and measured through a pres-

#### Pressure equivalents

	atm	kg/cm <sup>2</sup>	psi	in. Hg*	in. H <sub>2</sub> O†	mm Hg*	mb	μ Hg*
Atmospheres	1	1.033	14.70	29.92	406.8	760	1013	
Kilograms per square centimeter	0.9678	1	14.22	28.96	393.7	735.6	980.7	
Pounds per square inch	0.0680	0.0703	1	2.036	27.68	51.71	68.95	
Inches of mercury*	0.0334	0.0345	0.491	1	13.60	25.40	33.86	
Inches of water†	0.00246	0.00254	0.0361	0.0736	1	1.868	2.486	
Millimeters of mercury*	0.00132	0.00136	0.0193	0.0394	0.535	1	1.333	1000
Millibars		0.00102	0.0145	0.0295	0.401	0.750	1	750.1
Microns of mercury*						0.001	0.00133	1

\* At 32°F (0°C); to convert to Hg at 60°F, multiply by 1.00283.

† At 39.2°F (4°C); to convert to H<sub>2</sub>O at 60°F, multiply by 1.00096.



sure-tight bearing or seal. The accuracy of a bell-type gage is about 1% of full scale.

**Expansible metallic-element gages.** These are in wide use throughout industry, due to their low cost and freedom from the operational limitations of liquid gages.

There are three classes: bourdon, diaphragm, and bellows. All forms—as single elements—are affected by variations in external (atmospheric) pressure and hence are generally used as gage elements. Accuracies vary from 0.1–2.0% of full scale, depending on materials, design, and precision of components.

These elements may be designed to produce either motion or force under applied pressure. The more common motion type may directly position the pointer of a concentric indicating gage; position a linkage to operate a recording pen, or pneumatic relaying system to convert the measurement into a pneumatic signal; or position an electrical transducer to convert to an electrical signal.

**Bourdon-spring gages,** in which pressure acts on a shaped, flattened, elastic tube, are by far the most widely used type of instrument for pressures from 15 to 100,000 psi. These gages are simple, rugged, and inexpensive. See **BOURDON-SPRING PRESSURE GAGE**.

In diaphragm-element gages, pressure applied to one or more contoured diaphragm disks acts

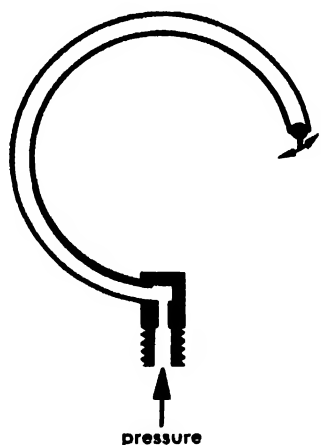


Fig. 3. Bourdon tube.

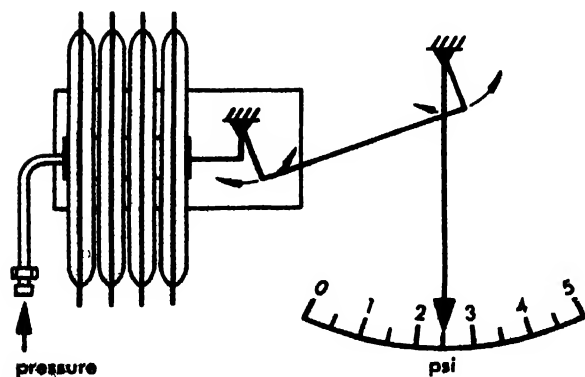


Fig. 4. Diaphragm-element gage.

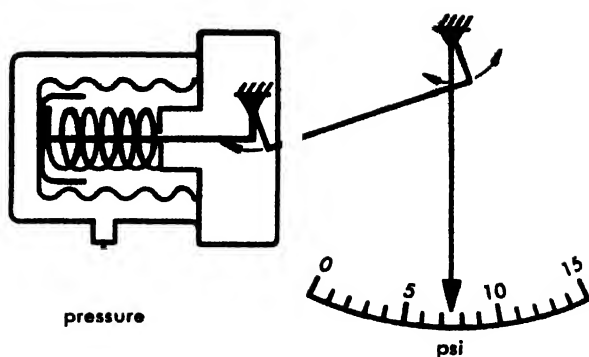


Fig. 5. Bellows-element gage.

against a spring or against the spring rate of the diaphragms, producing a measurable motion. The size, number, and thickness of the disks determine the range, from 2 in. of water to 30 psig.

For lower pressures, slack membrane diaphragms are used, opposed by a calibration spring; these instruments can detect differential pressures as low as 0.01 in. of water.

In bellows-element gages, pressure in or around the bellows moves the end-plate of the bellows against a calibrated spring, producing a measurable motion.

**Electrical pressure transducers.** These devices convert a pressure to an electrical signal. They are used for applications requiring unusually high speed of response, high sensitivity, extremely high or low pressure measurement, or for applications where an electrical signal representing pressure is more convenient to use than a mechanical motion representing pressure. See **PRESSURE TRANSDUCER**; **STRAIN GAGE**.

**Measurement standards.** For pressures below 20 psig, the universally accepted standard of pressure measurement—both in the laboratory and in the industrial plant—is the classic manometer, using mercury or water.

For higher pressures, the standard is the dead-weight tester. The principle is the balance of the force exerted by a precisely known weight on a piston of precisely measured area against a variable hydraulic pressure.

The pressure gage which is to be checked or calibrated is connected to the hydraulic reservoir. Weights corresponding to the check pressure are placed on the piston. The piston and weights are rotated to reduce the effect of friction. Hydraulic pressure is increased. When the desired check pressure is reached, the measuring piston floats freely.

The dead-weight tester is often used as a primary standard for laboratory use, and secondary standards (test gages which have been calibrated against the dead-weight tester) are used as field or plant test instruments. Dead-weight equipment is produced having accuracy to within  $\frac{1}{10}$  of 1% of reading. See **INSTRUMENTATION**; **PHYSICAL MEASUREMENT**. [B.D.H.; H.C.P.]

**Bibliography:** D. M. Considine (ed.), *Process Instruments and Controls Handbook*, 1957.

## Pressure transducer

An instrument component which detects a fluid pressure and produces an electrical signal related to the pressure. See TRANSDUCER.

In general, the complete instrument system comprises a pressure-sensing element, such as a bourdon tube, bellows, or diaphragm element; a device which converts motion or force produced by the sensing element to a change of an electrical parameter; and an indicating or recording instrument. Frequently the instrument is used in an automatic control loop to maintain a desired pressure. See PRESSURE CONTROL, AUTOMATIC.

Electrical measurement of pressure may be preferred to mechanical or pneumatic measurement for a number of reasons: (1) Electricity permits transmission of signals over longer distances. (2) Sometimes electricity permits more accurate or more rapid measurement. (3) It is usually easier to switch an electrical signal. (4) Electricity also may be economically or functionally most convenient.

Pressure transducers may be classified by the operating principle as resistive transducers, strain gages, magnetic transducers, crystal transducers, and capacitive transducers.

**Resistive pressure transducers.** Pressure is measured by these transducers by an element that changes its electrical resistance as a function of pressure.

Most types of resistive pressure transducers use a movable contact, positioned by the pressure-sensing element. The most common form is a contact sliding along a continuous resistor, which may be straight wire, wire-wound, or nonmetal such as carbon. If the cross section of the resistor is constant, the change in resistance will be proportional to the motion of the contact. The cross section may be made nonuniform to give a nonlinear relation between motion and change of resistance.

The resistance element may be curved or part of an arc for convenience in measuring angular motion.

Figure 1 shows one type of resistive pressure transducer. A bellows opposed by a precisely designed spring senses the pressure and converts the pressure to a linear motion of the plate between the bellows and the spring. The plate bears a con-

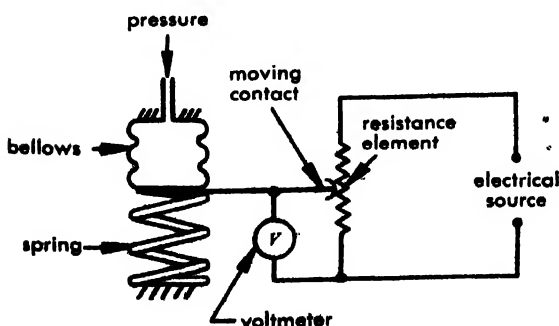


Fig. 1. Resistive pressure transducer.

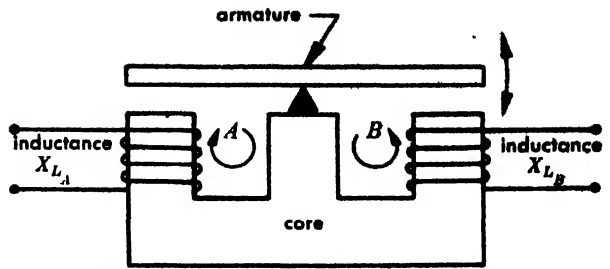


Fig. 2. Reluctance-type pressure transducer.

tact which wipes the surface of the precision wire-wound resistor. If a constant potential (ac or dc) is maintained across the resistor and if the resistance of the voltmeter is high with respect to the resistor, the measured voltage is a precise measure of the pressure. Instruments of this type are available with ranges as low as 0-5 pounds per square inch (psi) and as high as 0-10,000 psi. Accuracy of  $\frac{1}{4}\%$  of full scale is attainable.

The carbon pile may be considered to be another form of resistive pressure transducer. The simplest form is the carbon microphone, consisting of a box of carbon particles covered with a diaphragm. Varying pressure on the diaphragm varies the contact area of the carbon particles, hence the electrical resistance across the box. The carbon microphone is used only for dynamic measurements (changes) and has no calibrated accuracy. A modification, using a stack of processed disks in place of the carbon particles, is much more stable. See MICROPHONE.

Pressure-sensitive wire is yet another way of measuring pressure. When fluid pressure is applied to a wire, the wire is compressed and the resistance increases. Gold-chromium and manganin wire are ordinarily used because they have unusually low temperature coefficients of resistance. Pressures as high as 200,000 psi can be measured with these transducers.

**Strain-gage pressure transducers.** These might be considered to be resistive transducers, but are usually classified separately.

Strain-gage pressure transducers convert a physical displacement into an electrical signal by use of the fact that when a wire is stretched, its diameter is decreased and its electrical resistance is increased. The change in resistance is a measure of the displacement, hence of the pressure. Its advantages include infinite resolution and small size. The strain gage is usually used in a Wheatstone-type bridge circuit. For a more complete discussion, see STRAIN GAGE.

**Magnetic pressure transducers.** In this type, a change of pressure is converted into change of magnetic reluctance or inductance when one part of a magnetic circuit is moved by a pressure-sensing element—bourdon tube, bellows, or diaphragm. Magnetic transducers may attain an accuracy of 0.1% of full scale.

**Reluctance-type pressure transducer.** This type produces in a magnetic circuit a change of mag-

netic reluctance which is directly related to pressure. The change of reluctance is usually within one or two coils, wound intimately about the magnetic material in the magnetic circuit.

A representative ratio-type reluctance-changing device is shown in Fig. 2. A bourdon tube or other pressure-sensing device rotates the armature.

The reluctances (difficulty of passage of magnetic flux) in the magnetic paths A and B are determined chiefly by the lengths of the air gaps between the armature and the core. The inductance and inductive reactance of each winding depend on the reluctance in its magnetic path.

If the armature is at a neutral symmetrical position, the air gaps are equal, and the inductive reactances  $X_{LA}$  and  $X_{LB}$  are equal. Change of pressure decreases one air gap and increases the other, thus changing the ratio of the inductive reactances  $X_{LA}$  and  $X_{LB}$ . These changes can be used in a variety of circuits to produce an electrical signal which is a measure of pressure. The signal is transmitted to a measuring or controlling instrument.

**Inductive-type pressure transducer.** A change in inductance and inductive reactance of one or more windings is produced by the movement of a magnetic core that is positioned by a bourdon tube or other pressure-sensing element. Unlike the action of a reluctance-type transducer, the inductance change is caused by a change in air gap within the winding, rather than in a relatively remote portion of the magnetic circuit.

Figure 3 shows a representative ratio-type inductive device. The pressure-sensing element moves the core in response to changes of pressure. When the core is in a central position, the inductances of the two coils are equal. When pressure change moves the core, the ratio of the two inductances is changed. Energy is supplied to the coils by the same bridge circuit that measures the ratio of inductances. See INDUCTANCE BRIDGE.

Figure 4 shows another form of inductive pressure transducer, a differential transformer. A change of pressure moves the core, changing the coupling and the ratio of inductance of the two secondary windings.

When the core is centered, equal voltages are induced in the two oppositely wound windings, and the output voltage is zero. Change of pressure moves the core, increasing the voltage induced in one secondary and decreasing the voltage induced in the other. The change in output (differential) voltage is thus a measure of the pressure.

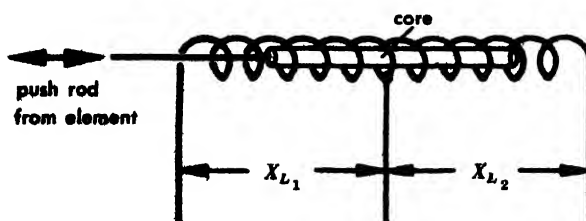


Fig. 3. Inductive pressure transducer.

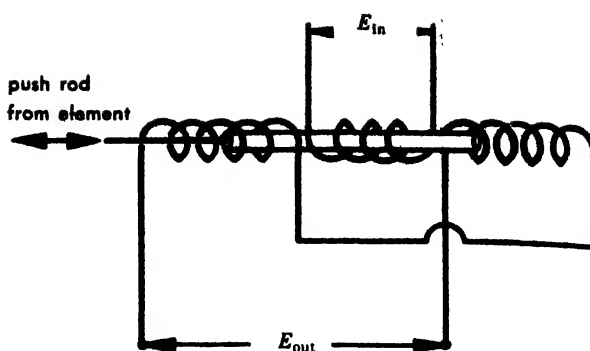


Fig. 4. Differential transformer.

**Crystal pressure transducers.** A crystal produces an electric potential when placed under stress by a pressure-sensing element. The stress must be carefully oriented with respect to a specific axis of the crystal.

Suitable crystals include naturally occurring quartz and tourmaline, and synthetic crystals such as Rochelle salts and barium titanate. The natural crystals are more rugged and less subject to drift of calibration. Although the synthetic crystals offer much higher voltage output, an amplifier is almost always required.

Crystal transducers offer a high speed of response, up to 1,000,000 cps. They are widely used for dynamic pressure measurements in such applications as ballistics and engine pressures. See PIEZOELECTRIC CRYSTAL.

**Capacitive pressure transducers.** Almost invariably, these sense pressure by means of a metallic diaphragm, which is also used as one plate of a capacitor. Any variation in pressure changes the distance between the diaphragm and the other plate, thereby changing the electrical capacitance of the system. The change in capacitance can be used to modify the amplitude or the frequency of an electrical signal.

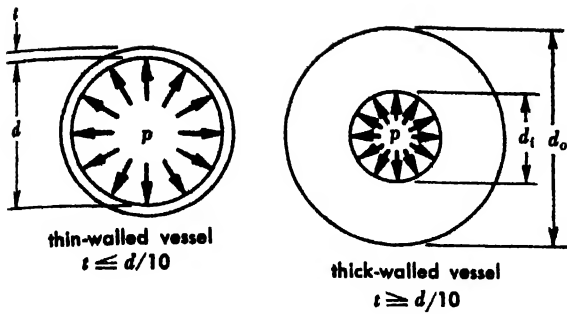
Transducers of this type are available to measure pressures as low as one-millionth of one psi and as high as 10,000 psi with accuracies of  $\frac{1}{4}\%$  attainable. See PRESSURE MEASUREMENT; TRANSDUCER, UNDERWATER. [B.D.H.; H.G.P.]

**Bibliography:** D. M. Considine (ed.), *Process Instruments and Controls Handbook*, 1957.

## Pressure vessel

A cylindrical or spherical metal container capable of withstanding pressures exerted by the material enclosed. Pressure vessels are important because many liquids and gases must be stored under high pressure. Special emphasis is placed upon the strength of the vessel to prevent explosions as a result of rupture, which would be dangerous to life and property. Codes for the safety of such vessels have been developed that specify the design of the container for specified conditions.

**Construction.** Most pressure vessels are required to carry only low pressures and thus are constructed of tubes and sheets rolled to form cylin-



Pressure vessels for moderate and for high pressures.

ders. Some pressure vessels must carry high pressures, however, and the thickness of the vessel walls must increase to allow adequate strength. Hydraulic and pneumatic cylinders are machine elements that are forms of pressure vessels.

Fabrication methods depend upon the vessel diameter, its wall thickness, and the type of cylinder ends employed. For extreme strength, heavy forgings may be welded together; for most normal vessels, rolled sheet and formed ends are fastened together with rivets.

Shell stress in pressure vessels is further dependent upon the type of end construction of the cylinder or vessel, whether welded, riveted or cast; the kind of material, whether ductile or brittle; and the conditions of operation, including pressure and temperature variations and limitations. In choosing the safe allowable stress, these variables are considered.

Although most pressure vessels contain an internal pressure, occasionally an external pressure is applied, which if excessive could buckle the sides and ends of the vessel. Such conditions depend on the elasticity of the material and result in buckling under critical pressures in a manner similar to the critical loads on columns. See COLUMN.

**Design.** Thin-walled pressure vessels are assumed to have uniform stresses through the wall thickness  $t$  if the diameter  $d$  is 10 or more times as great. For a pressure  $p$ , the shell stresses  $s_t$  are maximum in the circumferential direction, and this stress has the value  $s_t = pd/2t$ . In the axial direction, the stresses are half as great.

Thick-walled pressure vessels have a hyperbolic stress distribution through the wall thickness if the diameter is less than 10 times the thickness, with the maximum stress  $s_t$  at the inside surface. The stress is

$$s_t(\max) = \frac{p(d_o^2 + d_i^2)}{(d_o^2 - d_i^2)}$$

where  $d_o$  is the outside and  $d_i$  is the inside diameter. See HIGH-PRESSURE PROCESSES. [J.J.R.]

## Prestressed concrete

Concrete with stresses induced in it before use so as to counteract stresses that will be produced by loads.

Prestress is most effective with concrete, which is weak in tension, when the stresses induced are compressive. One way to produce compressive prestress is to place a concrete member between two abutments, with jacks between its ends and the abutments, and to apply pressure with the jacks. Another way, by far the most common, is to stretch steel bars or wires, called tendons, and anchor them to the concrete; when they try to regain their initial length and the concrete resists, they prestress the concrete. The tendons may be stretched with jacks or by heating (electrically).

Prestressed concrete is particularly advantageous for beams. It permits steel to be used at stresses several times larger than those permitted for reinforcing bars. It permits high-strength concrete to be used economically, for in designing a member with reinforced concrete all concrete below the neutral axis is considered to be in tension and cracked, and therefore ineffective, whereas the full cross section of a prestressed concrete beam is effective in bending. See CONCRETE BEAM.

An especially desirable characteristic of prestressed concrete is that as long as the material is maintained in compression it cannot crack. If cracks should appear under small overload, they generally will close when the load is removed. Sometimes concrete is prestressed principally to prevent cracking.

**Basic principles.** The effect of compressive prestress may be likened to picking up a group of books by applying pressure to the end pair. As long as the pressure is large enough, none of the books will slip out.

If this concept were applied to a concrete beam in actual practice, steel tendons would be tensioned and placed along the centroidal axis of the beam. The resulting prestress would result in a uniform compression at every section (Fig. 1). Loads would produce both tensile and compressive stresses at the middle of the span. The prestress would combine with these to increase the compression and cancel out the tension. The whole concrete section would be effective in resisting bending, and there would be no cracks.

In practice, however, tendons are rarely placed along the centroidal axis. A smaller prestressing force is required, and therefore less steel for the tendons, if the steel is placed below the centroidal

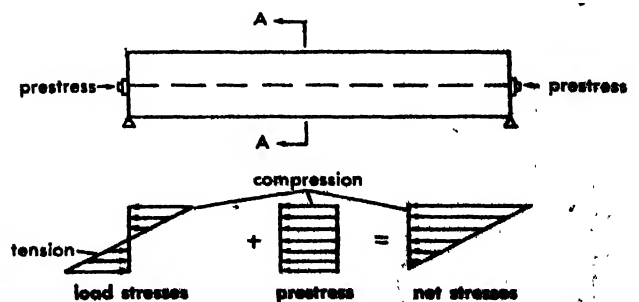


Fig. 1. Stresses at section A-A of a beam with uniform prestress.

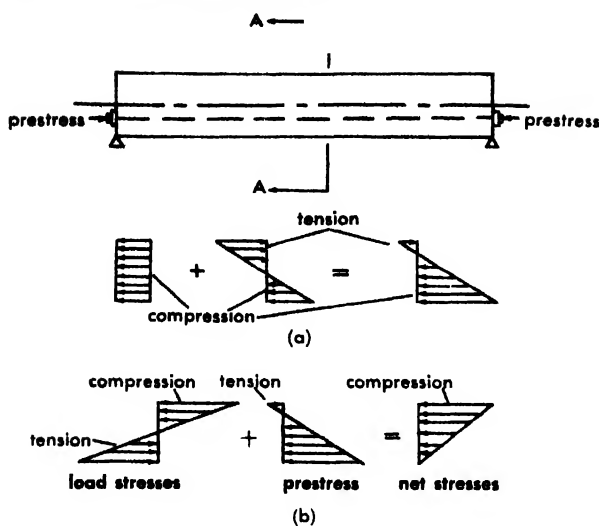


Fig. 2. Stresses at section A-A of a beam with eccentric prestress. (a) In the unloaded beam the simple-bending stress component is largely counteracted by the uniform compression component. (b) When a load is applied, all tension components are counteracted and only compression remains.

axis of the beam. With the eccentric prestress, stresses at each section of the unloaded beam may vary from tension at the top to compression at the bottom (Fig. 2).

When loads are applied to the beam, they produce both tensile and compressive stresses at the middle of the span. At the top of the beam they cause compressive stresses, which are reduced by the tensile prestress there. Elsewhere, the tensile stresses produced by the loads are counteracted by the compressive prestress.

With this arrangement of the tendons, there is a possibility that near the ends of the beam the tensile prestress may exceed the compressive stresses produced by the loads. The net tension may be undesirable, even though very small. To avoid this condition, the tendons may be draped in a vertical curve (Fig. 3). The distribution of prestress at any section of a beam so prestressed is similar to that for straight tendons applying an eccentric prestress except that the stresses decrease from midspan toward the ends, as do the bending stresses due to the loads. The draped arrangement of the tendons also is advantageous in counteracting diagonal tension near the ends of the beam.

Continuous beams may be prestressed in a similar manner. The tendons may be placed near the bottom of the beams near midspan and near the top over supports.

**Tendons.** Tendons generally are made of high-strength steel so that they can serve at high working stresses—for bars about 80,000 psi and for wires well over 100,000 psi, compared with the 20,000 psi or less ordinarily permitted for reinforcing steel.

Tendons must be tensioned to high stresses before being anchored to the concrete because losses

in stress due to shrinkage and plastic flow of the concrete are relatively high. If the tendons were tensioned to only 20,000 psi, for example, they might lose nearly all the prestress in a few months. But at 100,000-psi tension, the loss might be only about 15% because the increase in stress loss is not proportional to the increase in prestress.

Creep, or plastic flow, of concrete or steel is the inelastic deformation dependent on time and resulting solely from stress. Concrete shrinks when it dries and chemical changes take place; shrinkage is dependent on time but not on stresses due to external loading. In the absence of specific information, the loss of prestress due to shrinkage, when tendons are anchored to the concrete throughout the length of a member before shrinkage has taken place, may be assumed to be 0.0002 in./in., or 6000 psi for steel with a modulus of elasticity of 30,000,000. When tendons are anchored to the concrete after it has cured for several days, the prestress loss due to shrinkage may be about 3000 psi. The loss due to creep of the concrete may be assumed as 2.25 times the elastic compression. See CREEP OF MATERIALS.

Other possible losses in prestress that should be considered include those due to creep of the steel and to friction when the tendons rub against the concrete. See PLASTIC DEFORMATION OF METAL.

Wires are used for prestressing much more frequently than bars because of their greater strength. They may be used singly, in pairs, in cables composed of several parallel wires, or in strands. They may be stretched by electric heating, but by far the most common method of tensioning is with jacks. Various devices are used for gripping or anchoring tendons, including swaged fittings on strands, threads on bars, wedges, and buttonheads on wires.

**Pretensioning and posttensioning.** Two methods are used in fabricating prestressed beams. In one method, the concrete is bonded to the stretched steel before the prestress is applied. This is called pretensioning. In the other method, posttensioning, the prestress is applied initially through end anchorages and the concrete may or may not be bonded to the steel.

In pretensioning, the steel is laid through the beam forms and stretched between external abutments. Next, concrete is placed in the forms and allowed to set. When it has gained sufficient strength, the external pull on the tendons is relieved, transferring the prestress to the concrete through bond. Suitable for mass production, this method can be used on casting beds several hundred feet long to produce many beams simultaneously.

In posttensioning, the tendons are prevented from bonding initially to the concrete, usually by

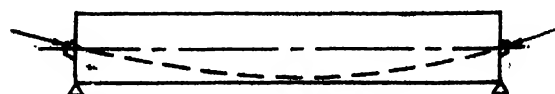


Fig. 3. Beam with tendons draped in a vertical curve.

encasement in sheaths. The concrete is placed in the beam forms around the sheathed tendons and allowed to set. When it has gained sufficient strength, jacks are used to tension the tendons and in so doing the jacks react against the ends of the beam. The tendons then are anchored to the concrete to apply the prestress.

Frequently, grout is forced into the sheaths to establish bond with the beam concrete. This gives the prestressed beam greater reserve strength and better crack control under overload. Posttensioning appears to be most advantageous for long-span beams and for assembling precast beam components in the field.

**Circular prestress.** Circular tanks, pipe, or the ring girder of domes may be prestressed, in contrast to the linear prestressing used for beams, by wrapping with steel bars or wires under high tension (see TANK). Special machines have been developed for rapid circular prestressing with wire. See REINFORCED CONCRETE. [F.S.M.]

**Bibliography:** Y. Guyon, *Prestressed Concrete*, 1953; A. Komendant, *Prestressed Concrete Structures*, 1952; K. Preston, *Prestressed Concrete*, 1960.

## Priapulida

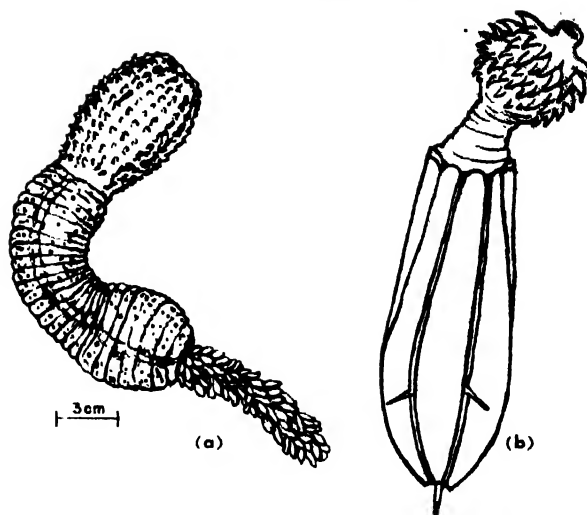
One of the minor groups of wormlike animals. Once linked with the Echiurida and the Sipunculida under the term "Gephyrea," it is now regarded as a separate phylum of the animal kingdom, with uncertain zoological affinities.

Priapulida are marine animals inhabiting the colder waters of both hemispheres. They live in burrows lined with mucus or in the sand or mud of the sea floor, and range from the intertidal area to waters of moderate depths.

The phylum is a small one with only two genera, *Priapulus* and *Halicryptus*. In the former there are four known species: *P. caudatus* and *P. bicaudatus* which occur in the northern temperate zone and Arctic waters, *P. tuberculato-spinosus* from Antarctic seas, and *P. horridus* from the coast of Uruguay. There is but one species of *Halicryptus*, *H. spinosus*, found in northern waters. *P. caudatus* and *P. tuberculato-spinosus* are very closely allied and have been considered by some authors to be the same species and to illustrate the phenomenon of bipolarity, the name given to the occurrence of the same or similar species in Arctic and Antarctic waters. With the exception of *P. horridus*, Priapulida are fairly common in their native habitat.

They are small to medium-sized animals, the largest specimens being only some 6 in. in length. The body is cylindrical and the trunk annulated but not segmented. The body of *Priapulus* is made up of three distinct portions, the proboscis, trunk, and caudal appendage. This latter consists of one or two stems thickly beset with tubules. *Halicryptus* has no caudal appendage.

The body wall is thick and muscular in all species and bears numerous spines and papillae. There are 25 rows of spines on the surface of the proboscis leading to the mouth, which is sur-



Priapulida. (a) *Priapulus*. To 6 in. in size. (b) *Priapulus* larva.

rounded by a number of comblike teeth. The pharynx is retractile and itself is beset with teeth. It is controlled by a series of retractor muscles attached to the inner wall of the trunk. The larva also possesses a retractile proboscis armed with teeth and is protected by chitinous shields. The sexes are usually separate. See ANIMAL KINGDOM. [A.C.S.]

**Bibliography:** F. Baltzer, *Echiurida*, in W. Kükenthal (T. Krumbach, ed.), *Handbuch der Zoologie*, vol. 2, 1931; K. Lang, Über die Entwicklung von *Priapulus caudatus* Lam., *Kgl. Fysiograf. Sällskap. Lund, Förh.*, 9(7):80-87, 1939.

## Prilling

A combination spray-drying and crystallizing technique used to produce agglomerates (prills) of ammonium nitrate for fertilizer. In prilling, a hot concentrated solution of ammonium nitrate is sprayed into a tower in which it crystallizes as it descends through a rising current of atmospheric air. See CRYSTALLIZATION; DRYING.

## Primary battery

An electric battery designed to deliver only one continuous or intermittent discharge. It cannot be recharged efficiently. Primary batteries are designed to deliver limited amounts of electric energy, determined by the materials used and the size of the cell. When the available energy drops to zero, the battery is usually discarded. Primary batteries may be classified by the type of electrolyte used.

**Aqueous-electrolyte batteries.** These batteries are solutions of acids, bases or salts in water as the electrolyte. These solutions have ionic conductivities of the order of 1 mho/cm and practically no electronic conductivity. Practical cells, such as the common Leclanche dry cell and the mercury cell, use aqueous electrolytes. They have the disadvantages of being corrosive to the electrode materials, having a relatively high evaporation rate of water





Fig. 1. Solid-electrolyte cell with solid crystalline salt electrolyte.

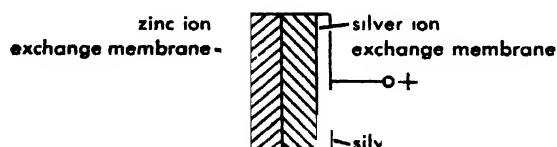


Fig. 2. Ion exchange bimetallic cell with solid electrolyte.

vapor which can cause cell failure, and being difficult to make free from leakage. For examples of cells with aqueous electrolytes see DRY CELL; MERCURY BATTERY; WET CELL.

**Solid-electrolyte batteries.** These use electrolytes of solid crystalline salts which have predominantly ionic conductivity. The conductivity is small compared with aqueous electrolytes and the current output is of the order of  $10^{-7}$  amp/in.<sup>2</sup>

Solid-electrolyte batteries may be classified in two broad categories: (1) cells with solid crystalline salt, such as silver iodide, as the electrolyte; (2) cells with ion-exchange membrane as the electrolyte. In either category, the conductivity must be nearly 100% ionic. Any electronic conductivity causes a continuous discharge of the cell and will limit the stand or shelf life.

A typical cell with solid crystalline salt electrolyte is the lead-lead chloride-silver chloride cell in Fig. 1. Here lead is the anode, lead chloride is the electrolyte, and silver chloride is the cathode. This cell has a potential of 0.49 volt. During discharge, lead is oxidized to lead ion and silver chloride is reduced to silver.

Cells with solid salt electrolyte have been developed into miniature batteries. One type delivers 90–100 volts at  $10^{-11}$  amp, and has a capacity of 1 amp-sec. This is over  $10^6$  days at  $10^{-11}$  amp. The practical life of the cell is much less but may be as much as 10 years at room temperature. It can be stored at 160°F for at least 30 days and will operate over the range  $-65$  to  $+165$ °F. The battery is  $\frac{3}{8}$  in. in diameter, 1 in. in length.

An example of a cell with ion exchange membrane as electrolyte is the zinc-zinc ion exchange membrane, silver ion exchange membrane-silver cell shown in Fig. 2. Physically, the metal electrodes are in contact with the solid membrane which contains two regions. The region adjacent to the zinc is in the zinc ion state. The region adjacent to the silver is in the silver ion state. The discharge reaction increases the zinc ion quantity and decreases the silver ion quantity, in proportion to the amount of charge transferred. This cell has a potential of about 1.5 volts.

The zinc-silver cell described has serious shortcomings. The stand life is poor, indicating internal self-discharge, and the capacity is limited by the available supply of silver ions. In strongly ionized types of ion exchange material, the volume density of ionizing sites is about 1 equiv/liter, or 0.4 amp-hr/in.<sup>3</sup> This is very low compared with metal-oxide cathodes.

A cell with higher capacity can be made by replacing the silver ion exchange material and silver by manganese dioxide plated on an inert metal, such as tantalum. This gives a capacity of about 100 times as much, for equal volume.

Ion exchange electrolytes are also used with hydrogen and oxygen gas electrodes. The electrodes consist of platinized metal screens. The electrolyte is a hydrogen ion exchange material. The room temperature emf of this cell is 0.96 volt. See ION-PERMEABLE MEMBRANE.

**Waxy-electrolyte batteries.** These use waxy materials, such as polyethylene glycol, in which a small amount of a salt is dissolved in the molten wax. At room temperatures these materials are solid. The conductivity is small and the current output is limited to about  $10^{-6}$  amp/in.<sup>2</sup>

Figure 4 shows a battery stack of cells using a waxy electrolyte. The electrodes are sheet zinc and manganese dioxide. The electrolyte is made of polyethylene glycol in which is dissolved a small amount of zinc chloride. This electrolyte is melted and painted on a paper sheet to form the separator in Fig. 4.

A 25-cell stack, built as shown in Fig. 4 and measuring 0.34 in. in length and 0.25 in. in diameter, weighed 1.5 g. A 0.50 in.-diameter stack weighed 6.0 g. The initial open-circuit voltage was 37.5 volts (1.5 volts per cell).

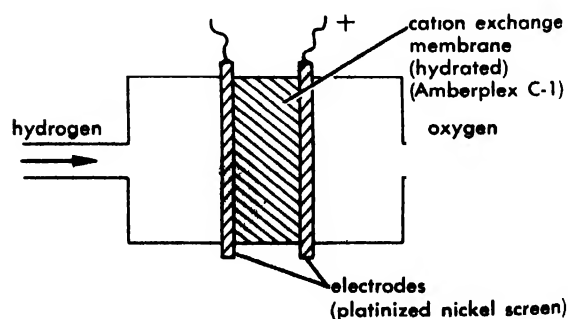


Fig. 3. Solid-electrolyte cell with ion exchange membrane as electrolyte.

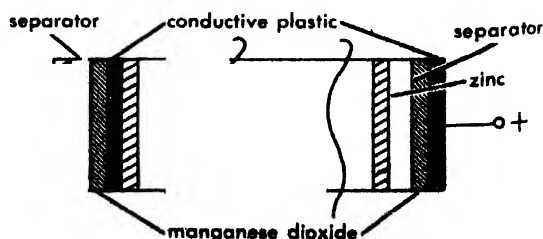


Fig. 4. Waxy-electrolyte battery stack.

The internal resistance of this cell is high, and increases as temperature decreases. This high internal resistance limits the usefulness of the cell, but it may be suitable for long-life potential sources of miniature size.

**Fused-electrolyte batteries.** These use crystalline salts or bases which are solid at room temperature. In use, the cell is heated and maintained at a temperature above the melting point of the electrolyte. See RESERVE BATTERY. [S.E.]

## Primates

An order of mammals that includes the tree shrews, tarsiers, monkeys, apes, and man. Although unquestionably a natural assemblage, the Primates are difficult to characterize, largely because they share no conspicuous morphological specialization comparable to those characterizing the bats, the whale, the carnivores, and most other orders of mammals. Such specializations are associated with the exploitation of particular ecological zones (for example, flying, swimming, and predation), and it is evident that the Primates are not so closely tied to an ecological zone as are most other orders of mammals. Throughout primate history there has been a trend toward adaptation for arboreal life, but perhaps most significant was the tendency to emphasize the brain.

**Fossil record of lower primates.** The most primitive living primates, the tree shrews, are known from the fossil record by a single specimen, *Anagale gobiensis*, from the early Oligocene of Mongolia. The tree-shrew stock apparently never reached Europe or the Americas, and living tree shrews are confined to southeastern Asia, the East Indies, and the Philippines.

Primates first appear in the fossil record in the Paleocene, and about 14 genera of lemurs are known from the Paleocene and Eocene of North America and Europe. All belong in the families Plesiadapidae and Adapidae, which disappear from the record before the end of the Eocene. The best known of these early lemurs is *Notharctus*, an animal about the size of a cat, of which several complete skeletons exist. *Notharctus* was very similar to the primitive existing lemurs of Madagascar. Lemurs disappeared from the Northern Hemisphere toward the end of the Eocene, but are abundantly represented in the living fauna of Madagascar.

Associated with these early lemurs was a second primate stock, the Anaptomorphidae, a group of tarsioids that underwent an extraordinary adaptive radiation during the Paleocene and Eocene. The sole survivor is the living tarsier of the East Indies and Philippines.

The lorises and galagos, now living in southeastern Asia and Africa, are practically unknown as fossils and their history is obscure. All are highly specialized nocturnal animals, well off the main line of primate evolution.

**Fossil record of higher primates.** Fossils of the higher primates are extremely rare, particularly from the Eocene and Oligocene when these animals

were presumably undergoing their major radiation. Consequently, relationships among the higher primates must be largely inferred from the anatomy and distribution of existing forms. It is certain that the ancestry of the South American primates has been distinct from that of the Old World primates at least since the early Eocene, but the ancestors of neither group are known. The living forms are divided into three superfamilies: the Ceboidea, Cercopithecoidea, and Hominoidea.

The Ceboidea (New World monkeys) are distinguished by having three premolars instead of two as in the other two groups. Two distinct families are represented, the Callitrichidae (squirrel monkeys), with claw-shaped nails and a nonopposable thumb, and the Cebidae with flat nails and an opposable thumb.

The Cercopithecoidea (Old World monkeys) include the baboons, guenons, and langurs of Asia and Africa.

The Hominoidea (apes and man) form a natural group united by a multitude of common features. The great apes (gorilla, orangutan, and chimpanzee) are obviously relict forms on the way to extinction. The gibbons, smaller and more widely distributed, are still flourishing. See EUTHERIA; PRIMATES (FOSSIL); ZOOGEOGRAPHY. [D.D.D.]

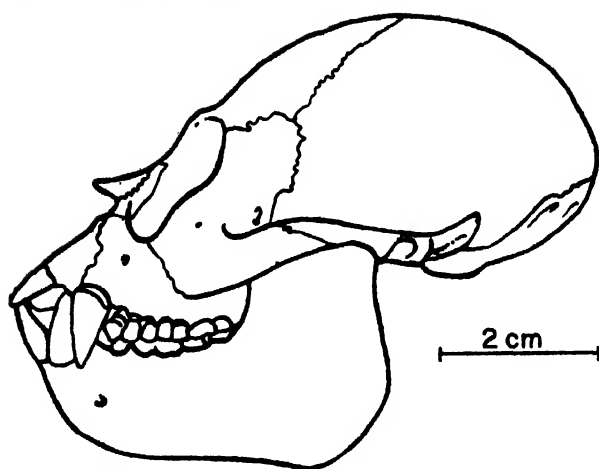
## Primates (fossil)

The order Primates usually includes the tupaoids, lemurids, lorises, galagos, tarsioids, monkeys, apes, and man. The tupaoids are sometimes placed with the menotyphlan insectivores. In evolution toward the higher primates, natural selection tended to favor those with larger brains, better sight, and more dexterous front feet that functioned as hands. The incisors are reduced to two in each of the jaws in the monkeys and apes. See PRIMATES.

Small prosimian primates were common in the middle latitudes of North America and Europe from the Paleocene until the middle Eocene. Apparently, faunal interchange between the two regions was greatly reduced or ceased during the middle Eocene. Then as aridity increased in the late Eocene and early Oligocene, the distribution of the primates shifted southward in both hemispheres. Some of these prosimians were extremely close to primitive stocks of other orders, such as the Insectivora and the Dermoptera. Even in the groups that were surely primates, infraordinal affinities are obscured by the presence of both lemurid and tarsioid characters.

Primates, both Recent and middle and late Cenozoic, are the results of extensive diversification and evolution in little or unknown populations that tended to favor milder climates in lower latitudes.

True lemurs are confined to Madagascar, where they occur in late Pleistocene deposits and as living animals. The only fossil of a loris is one tooth from the Pliocene of India. Some excellent galago fossils have been found in a Miocene fauna in Kenya, eastern Africa. It has been suggested that these are ancestral to the living galagos. Although



Skull of *Cebupithecus sarmientoi*, a late Miocene New World monkey from Colombia, South America. (After R. A. Stirton, 1951)

there are numerous Eocene prosimians with tarsoid characters, there are no middle or late Tertiary forms linking them with the living tarsier.

New World monkeys (Cebidae) appeared for the first time in the late Oligocene and early Miocene of Argentina. They were already typical ceboids at that time. Thus, no clue is provided concerning their ancestral relationships. Others have been found in the late Miocene of Colombia. *Homunculus* is related to the howler monkeys, *Neosaimiri* to the squirrel monkeys, and *Cebupithecus* to the sakis and uakaris monkeys. An aberrant extinct genus, *Xenothrix*, has been discovered in a kitchen midden in Jamaica. It seems likely that the Cebidae arose from an early prosimian group from North America.

Old World monkeys (Cercopithecidae) and apes (Pongidae) seem to have descended from different groups of Eocene prosimians. The earliest cercopithecoid, *Moeripithecus*, occurs in the early Oligocene of Egypt. Other more advanced genera such as *Ankarapithecus*, *Mesopithecus*, and *Dolichopithecus* have been found in Pliocene faunas, mostly in central and southern Europe. Some of the living genera also extend back into the Pliocene. Baboons, recognized at a glance because of their elongated faces, first appeared in the Pliocene and are one of the most common cercopithecoid groups. One Pleistocene genus, *Dinopithecus*, rivaled *Megaladapis* (Lemuridae) and *Gigantopithecus* (Pongidae) as one of the largest primates.

Among the most interesting primates of the Old World is *Oreopithecus* from Mount Bamboli, Italy. Some of its characters have been emphasized in support of possible hominid relationships but allocation to the Hominoidea has not yet been firmly established.

There are many divergent lineages of apes and apelike creatures, all restricted to the Eastern Hemisphere. *Parapithecus*, known from a pair of lower jaws from the early Oligocene of Egypt with some prosimian characters, represents the most primitive of these groups. Other middle to late Ter-

tiary genera such as *Propliopithecus*, *Limnopithecus*, and *Pliopithecus*, although related to the living gibbons, may have been more terrestrial in habits. The orangutans, chimpanzees, and gorillas apparently are the descendants of dryopithecine apes as represented by the genera *Proconsul*, *Dryopithecus*, *Sivapithecus*, and others in Miocene and Pliocene faunas. These Tertiary genera also possess hominid characters indicating a common basic heritage with the Hominidae. Perhaps the most spectacular of all apes is the giant orangutanlike *Gigantopithecus* in the Pleistocene of China. See FOSSIL MAN. [M.C.MC.; R.A.S.T.]

## Prime mover

The component of a power plant that transforms energy from the thermal or the pressure form to the mechanical form. Mechanical energy may be in the form of a rotating or a reciprocating shaft, or a jet for thrust or propulsion. The prime mover is frequently called an engine or turbine and is represented by such machines as water wheels, hydraulic turbines, steam engines, steam turbines, windmills, gas turbines, internal combustion engines, and jet engines. These prime movers operate by either of two principles (Fig. 1): (1) balanced expansion, positive displacement, intermittent flow of a working fluid into and out of a piston and cylinder mechanism so that by pressure difference

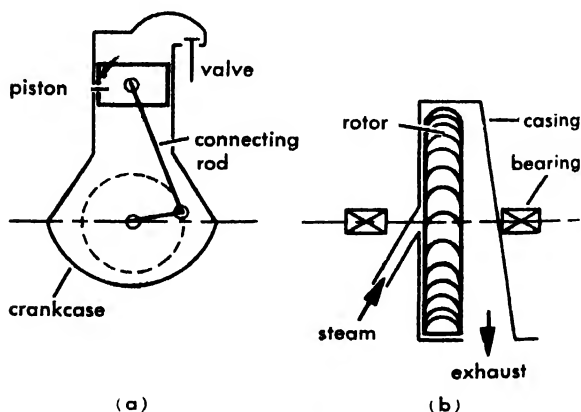


Fig. 1. Representative prime movers. (a) Single acting, four cycle, automotive type internal combustion engine. (b) Single stage, impulse type, steam turbine.

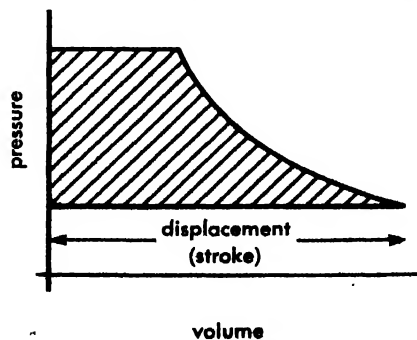


Fig. 2. Pressure-volume diagram (indicator card) for fluid-displacement type of prime mover.

Table 1. Dimensional and performance criteria of some selected fluid displacement type prime movers

Type	Size, hp	Rpm	Stroke, in.	Bore: stroke ratio	Piston speed, ft/min	Brake mep, psi	Diagram factor, or engine efficiency
Steam engine	25-500	100-300	6-24	0.8-1.2	400-600	50-100	0.6-0.8*
Automobile engine	10-300	2000-4000	3-5	0.9-1.1	1000-2000	50-100	0.4-0.6†
Aircraft engine	100-3000	2500-3500	4-7	0.8-1.1	1500-3000	100-230	0.4-0.6†
Diesel, low-speed	100-5000	100-300	10-24	0.8-1.0	500-1000	40-80	0.4-0.7†
Diesel, high-speed	25-1000	1500-2000	3-6	0.8-1.0	800-1500	50-100	0.4-0.6†

\* Logarithmic standard.

† Air-card standard.

Table 2. Dimensional and performance criteria of some selected fluid acceleration type prime movers

Type	Rating, kw	Number of stages	Head, ft or pressure, psi	Temperature, °F	Exhaust pressure, in. Hg abs	Rpm	Tip speed, ft/sec	Efficiency
Pelton water wheel	1000-50,000	1	500-5000 ft	Ambient	atm	100-1200	100-250	0.75-0.85
Francis hydraulic turbine	1000-100,000	1	50-1000 ft	Ambient	atm*	72-360	50-200	0.8-0.9
Propeller (and Kaplan) hydraulic turbine	5000-100,000	1	20-100 ft	Ambient	atm*	72-180	70-150	0.8-0.9
Small condensing steam turbine	100-5000	1-12	100-400 psi	400-700	1-5	1800-10,000	200-800	0.5-0.8
Large condensing steam turbine	100,000-500,000	20-50	1400-5000 psi	900-1200	1-3	1800-3600	500-1500	0.8-0.9
Gas turbine	500-10,000	10-20	70-100 psi	1200-1400	atm	3600-10,000	500-1500	0.8-0.9

\* Draft tube gives negative pressure on discharge side of runner.

on the opposite sides of the piston, or its equivalent, there is relative motion of the machine parts; or (2) free continuous flow through a nozzle where fluid acceleration in a jet (and vane) mechanism gives relative motion to the machine parts by impulse, reaction, or both.

**Displacement prime mover.** Power output of a fluid displacement prime mover is conveniently determined by pressure-volume measurement recorded on an indicator card (Fig. 2). The area of the indicator card divided by its length is the mean effective pressure (mep) in psi, and horsepower of the prime mover is given by

$$\text{horsepower} = \frac{\text{mep} \times L a n}{33,000} \quad (1)$$

where  $L$  is stroke in feet,  $a$  is piston area in square inches, and  $n$  is number of cycles completed per minute. Actual mep is smaller than the theoretical mep and is related to the theoretical value by diagram factor or engine efficiency (Table 1).

**Acceleration prime mover.** Performance of fluid acceleration (hydraulic) prime movers is given by the equation

$$\text{horsepower} = \frac{QH}{8.8} \times \text{efficiency} \quad (2)$$

where  $Q$  is water flow rate in cubic feet per second, and  $H$  is head in feet. For heat-power prime movers of the fluid acceleration type, actual properties of the thermodynamic fluid, as given in tables and graphs, especially the Mollier chart, permit the rapid evaluation of the work or power output from the general energy equation which resolves to the form

$$\Delta W, \text{ Btu/lb of fluid} = h_{\text{inlet}} - h_{\text{exhaust}} \quad (3)$$

where  $h$  is the enthalpy in Btu/lb, and the inlet and exhaust conditions can be connected by an isentropic expansion for ideal conditions, or modified for irreversibility to a lesser difference by engine efficiency (Fig. 3, Table 2). Fluid consumption follows from

$$\text{fluid consumption, lb per hphr} = 2545/\Delta W \quad (4)$$

or

$$\text{lb per kw hr} = 3413/\Delta W \quad (5)$$

In the fluid acceleration type of prime mover, jet velocities experienced in the nozzles can be found in feet per second, for nonexpansive fluids by

$$\text{jet velocity} = C \sqrt{2gH} = 8.02 C \sqrt{H} \quad (6)$$

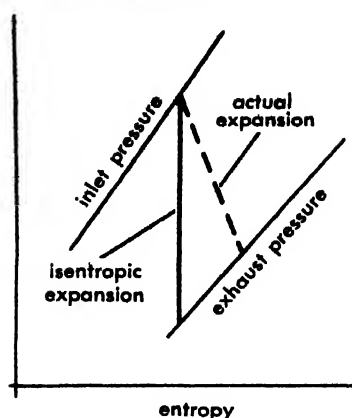


Fig. 3. Enthalpy-entropy (Mollier) chart showing performance of steam- or gas-turbine type of prime mover.

and, for expansive fluids, by

$$\text{jet velocity} = C \sqrt{2g\Delta W} = 223.7 C \sqrt{\Delta W} \quad (7)$$

where  $H$  and  $\Delta W$  are as given above and  $C$  is the velocity coefficient, seldom less than 0.95 and usually from 0.98 to 0.99.

Selected representative performance values of some prime movers are presented in Tables 1 and 2. For further details see INTERNAL COMBUSTION ENGINE; STEAM ENGINE; TURBINE. [T.B.]

**Bibliography:** A. H. Gibson, *Hydraulics and Its Applications*, 4th ed., 1930; L. C. Lichty, *Internal Combustion Engines*, 1951; T. Baumeister (ed.), *Marks' Mechanical Engineers' Handbook*, 6th ed., 1958; J. K. Salisbury, *Steam Turbines and Their Cycles*, 1950; E. T. Vincent, *The Theory and Design of Gas Turbines and Jet Engines*, 1950.

## Primer (explosive)

An agent used with explosives, propellants, and pyrotechnics to produce the initial fire. The primer itself may be initiated by percussion, stab action, friction, or heat. The gases and hot particles from the primer serve in turn to ignite a larger mass of material, for example, the gunpowder in a cartridge, the primary explosive in a detonator cap, or the powder train in a fuze. The term primer is sometimes loosely applied to a detonator.

The amount of primer used is usually about 0.1 gram or less. A common example is the tiny globeule that may be seen on the ends of the fuse wire post in a photographic flashbulb; this material serves to ignite the aluminum foil.

Although primer compositions deflagrate rapidly, they do not detonate. In fact, a detonating action would be most undesirable because it would tend to blow away rather than to ignite the charge. The total burning time of a primer is about 500 microseconds.

Primer mixtures sensitive primarily to friction may consist simply of potassium chlorate as the oxidizer and antimony sulfide as the fuel. Such a mixture produces hot particles as well as gases.

Ground glass or silicon carbide may be added to increase the frictional effect, and sulfur to make the mixture quicker (more sensitive). Also, very fine-grained black powder (meal powder) is sometimes added. The ingredients are made up in a paste with a small amount of gum arabic which binds them together when the paste dries.

Cartridge percussion caps usually contain a primary explosive such as mercury fulminate in addition to the materials mentioned above. However, primary explosives are not essential; they may be replaced by other materials such as lead thiocyanate. Lead compounds seem to be preferred, probably because of the hot particles of lead oxide that they produce. Lead styphnate (2,4,6-trinitrorescorcinol) is now a common percussion-cap component.

In detonator caps, the primer is often of a match-head composition. In fact, the formulation of primers is quite similar to that of matches and pyrotechnics; and ingredients such as potassium chlorate, sulfur, and antimony sulfide are common in all applications. See DETONATOR; EXPLOSION AND EXPLOSIVE; MATCH; PYROTECHNICS. [W.E.G.]

**Bibliography:** T. L. Davis, *Chemistry of Powder and Explosives*, 1943.

## Primer (surface coating)

A material used for the first coat of paint. Primers are designed to promote adhesion of the coating system to the substrate, to furnish a good base for further coatings, and to prevent attack on the substrate by air, water, or other materials. Primers are not intended to contribute exterior durability or appearance.

Primers for wood are formulated to give the maximum adhesion to the wood and to provide sufficient flexibility to adjust to the dimensional changes which occur when the wood swells and shrinks because of changes in moisture content. When designed for exterior wood they should also be resistant to the penetration of moisture and to the action of moisture in destroying adhesion. Primers for interior wood, often called undercoaters, must also have good adhesion to wood, and because they are often used under enamels, must give a smooth film and permit easy sanding.

Primers for metal must, in addition to giving good adhesion, prevent the spread of rust from scratches or other damage which uncovers the surface. For this purpose, they usually contain a corrosion-inhibiting pigment, which retards the electrochemical reactions that cause corrosion. The most common inhibitive pigments are red lead, zinc chromate, and basic lead chromate. A number of other related pigments are often used, and numerous others have been suggested. Because all these are somewhat toxic, primers based on red iron oxide are often used when they are to be dry sanded, or when, for other reasons, toxic materials must be avoided. These pigments do not have any inhibitive action but rely on a tight film to prevent corrosion.

Primers for galvanized iron usually contain substantial quantities of metallic zinc, because other types do not, in general, adhere well to galvanized iron, especially on exterior exposure. These primers may also be used on other metals, in which case the sacrificial corrosion of the zinc protects the underlying metal.

Primers for porous materials must seal the surface adequately to provide a uniform base for future coatings. Because many of these materials are alkaline in reaction, high alkali resistance is another important requirement. Such materials are often called sealers. Porous surfaces often aid in adhesion through mechanical interlocking, so that extreme adhesion, chemically, is not always required. Because many of these porous surfaces are rough, many primers are also designed to fill the surface and provide a smoother surface for future paints.

When the surface to be painted is attacked by the solvents in the topcoats, a primer which will seal the underlying surface and isolate it from attack must be used. This is a common problem in painting over asphalt or other bituminous materials. In this case a water- or alcohol-thinned primer is indicated. In other cases, the primer must be chosen for the particular problem.

Primers are always pigmented. In clear finishes the coat which performs this function is described as a sealer, an undercoater, or a wash coat. See CORROSION; PAINT; SURFACE COATING.

[F.S.D.]

## Primitive gut

The tubular structure in embryos which differentiates into the alimentary canal. The method by which the primitive gut arises depends chiefly on the yolk content of the egg.

**Two-layered blastoderms.** Eggs with small or moderate amounts of yolk usually develop into spherical blastulae which invaginate at the vegetative pole to form double-walled gastrulae. The invaginated sac extends in length to become the primitive gut. In some groups, such as the echinoderms, chaetognaths, chordates, and amphibians, its external opening, the blastopore, persists as the anus. The mouth forms as a new opening by contact of the opposite end of the gut with the skin which then perforates. In other groups, like the annelids, arthropods, and mollusks, the blastopore becomes the mouth, and a new anal opening is formed.

**Three-layered blastoderms.** Animals with more yolk than can be cleaved, as fish, reptiles, and birds, form flattened gastrulae consisting of three-layered blastoderms surmounting the yolk. Mammals also belong in this group, although the yolk has been lost secondarily in all except the monotremes. The head is formed by a folding of the blastoderm upon itself. The entodermal layer within the head fold becomes the pharynx. This foregut is extended by an anterior growth of the whole head and by the union of lateral entodermal folds at its posterior

boundary. In most forms, the hindgut arises by a similar folding in the opposite direction, the tail fold, at the posterior end of the blastoderm. The shark is an exception in which the hindgut is formed by the union of two lateral entodermal ridges beneath the notochord, much as the medullary plate folds into neural tube above the notochord. The midgut remains open until the entodermal layer of the blastoderm surrounds the yolk. The communication of the midgut with the yolk sac gradually narrows to an umbilicus as the foregut, hindgut, and lateral body folds press in from all sides, but the midgut floor is brought into position only with the absorption of the yolk sac. Mouth and anus form by contact of the pharynx and hindgut with the ectoderm at their respective ends of the body. Biochemical incompatibilities of ectoderm and entoderm cause such areas of contact to perforate and form the body orifices.

Teleost fish are an exception to the above account in that the true entoderm does not spread beneath the whole blastoderm, but is confined to the embryonic area. Along the axis of the body beneath the notochord, the entoderm condenses into a solid rod which then hollows out secondarily to form the primitive gut. See CLEAVAGE, EMBRYONIC; DEUTEROSTOMIA; GASTRULATION; OVUM; PROTOSTOMIA.

[H.L.H.]

## Primulales

A small order of the plant subclass Dicotyledoneae containing 3 families with 70 genera and 2100 species. These are of no economic importance except as ornamentals. Two characteristic features are the opposite stamens and the free central placentation. The five stamens are opposite the five petals instead of being alternate with them because the outer circle of stamens is abortive and represented by mere rudiments called staminodia. The axis of the flower projects into the ovary cavity as a central shaft to which the ovules are attached, a condition designated free central placentation. Here belong such familiar ornamentals as ardisia, plumbago, cyclamen, and the primroses. See DICOTYLEDONEAE; EMBRYOPHYTES; PLANT KINGDOM.

[P.D.S.]

## Printed circuit

A generic term applied to a method of fabricating electric circuits, consisting of conductors and electronic components (resistances, inductances, and capacitances), by any of several graphic art processes. Printed circuits make economical mass production possible, save space and weight, and increase reliability of electronic equipment. They are used in practically all types of electronic equipment, such as radio and TV sets; electrical wiring behind the dashboard in automobiles; guided-missile and airborne electronic equipment; computers; and industrial control equipment.

The rapid adoption of the graphic art processes by the electronics industry is a demonstration of the effectiveness of those processes in achieving cost reduction and equipment miniaturization.



Printed circuits are of interest to industry for the following reasons:

1. Printed circuits are the common denominator for almost all approaches to the mechanized fabrication of electronic equipment.
2. Use of printed circuits has greatly reduced the labor required for the wiring of an electronic circuit. This is specifically true for small electronic units used in airborne or guided-missile equipment.
3. Printed circuits can be manufactured more uniformly because the graphic art processes are mechanized.
4. Uniformity of printed circuits improves the quality of the product through simplification of quality control.
5. Printed circuitry has helped to minimize one major cause of unreliability in electronic equipment by permitting the use of dip-soldering processes. (In dip soldering, the joints between the electronic component and the conductor are exposed to molten solder and joined in one precisely controlled operation.)

The most commonly used printed-circuit processes may be divided roughly into three main groups listed in the order of greatest acceptance: (1) material-removal processes, (2) film-deposition processes, and (3) mold and die processes.

**Material-removal processes.** Of the material-removal processes, photoetching and stencil etching are probably the most widely used techniques. These are used primarily in the fabrication of printed wiring.

**Photoetching.** In photoetching, the etchant-resist pattern (conductor pattern) is formed photographically and is capable of providing fine definition of the conductor lines, such as those required for small commutators and switch contacts. A photosensitive film is applied to a copper foil, which is bonded to a plastic laminate such as paper base phenolic. A photographic negative of the circuit pattern (originally drawn several times full size) is superimposed on the sensitized film, and the film is exposed to ultraviolet light. This method is similar to the production of a photographic positive. After exposure the photosensitive film hardens. The plate is then placed in alcohol which dissolves the unexposed film from the copper foil. The exposed areas are left covered by the hardened film, which serves to protect the copper foil during the subsequent etching process. The uncovered copper is next dissolved in an acid or ferric chloride etchant bath. Finally the hardened film is dissolved from the exposed areas, leaving copper conductors in the original circuit pattern (see Fig. 1).

**Stencil etching.** In stencil etching the protective film forming the circuit pattern is applied by a printing process such as silk screening. The protective film, usually an enamel, is dried, and the exposed copper is etched as previously described.

**Film-deposition processes.** The mechanical application technique, stencil screening, and electroplating are the deposition processes most often

used for the fabrication of circuits on either plastic or ceramic-base materials. For components such as precision resistors, vacuum-deposition techniques are used, usually on ceramic or glass base material. Chemical or photochemical reduction techniques

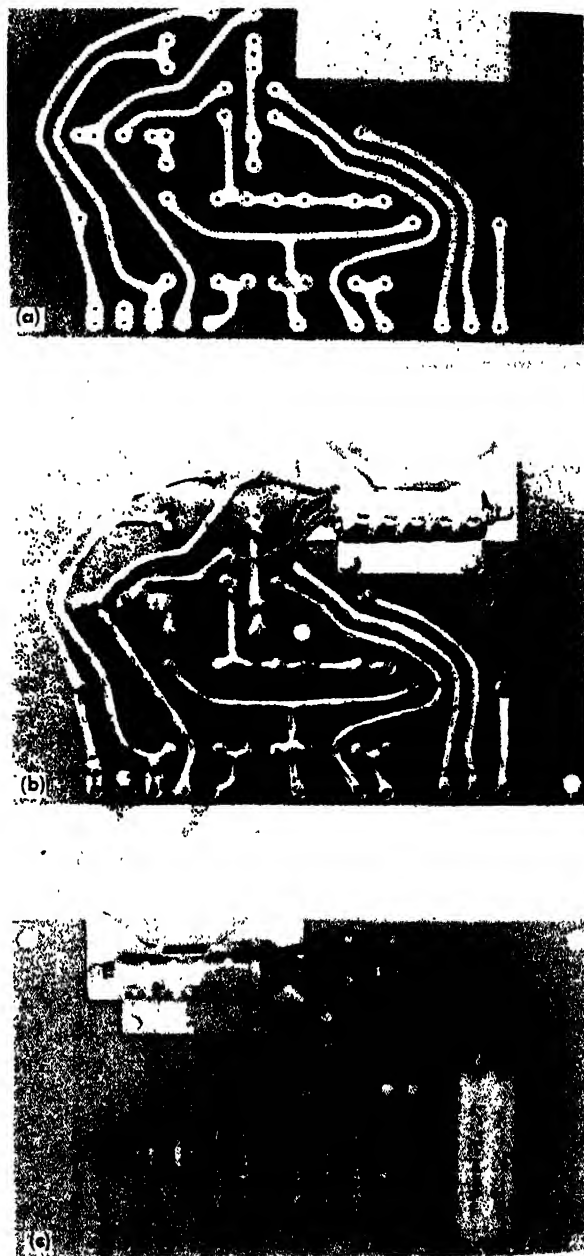


Fig. 1. Etched copper circuit on epoxy-glass laminate. (a) Copper circuit pattern with etchant resist removed is ready for drilling of holes to mount component parts. (b) Circuit pattern side of board with component parts mounted and lead connections soldered. Subminiature tube was mounted and tube connections made after the rest of the parts were dip-soldered. (c) Component side of board shows parts mounted through holes in circuit board. After the parts are soldered, a coating of epoxy resin is applied on both sides of the circuit board to serve as a moisture barrier and to bond the parts to the board. (Ramo-Wooldridge Corp.)

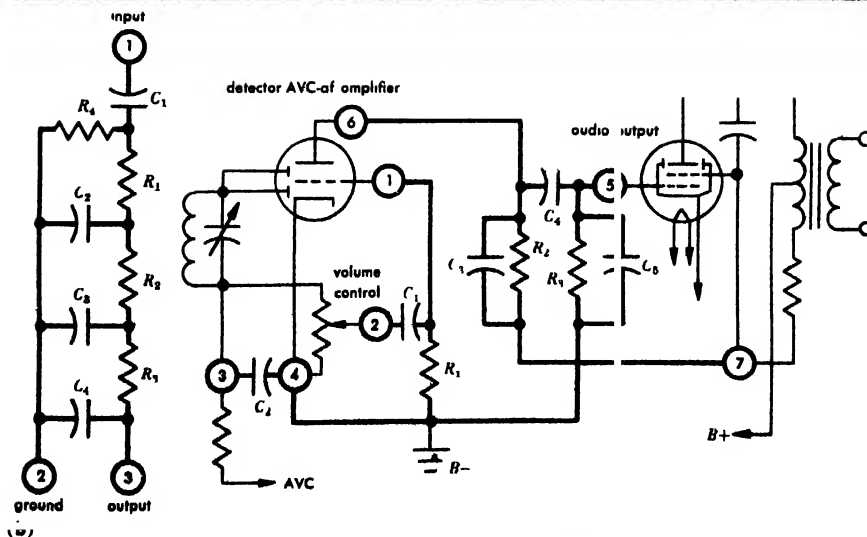
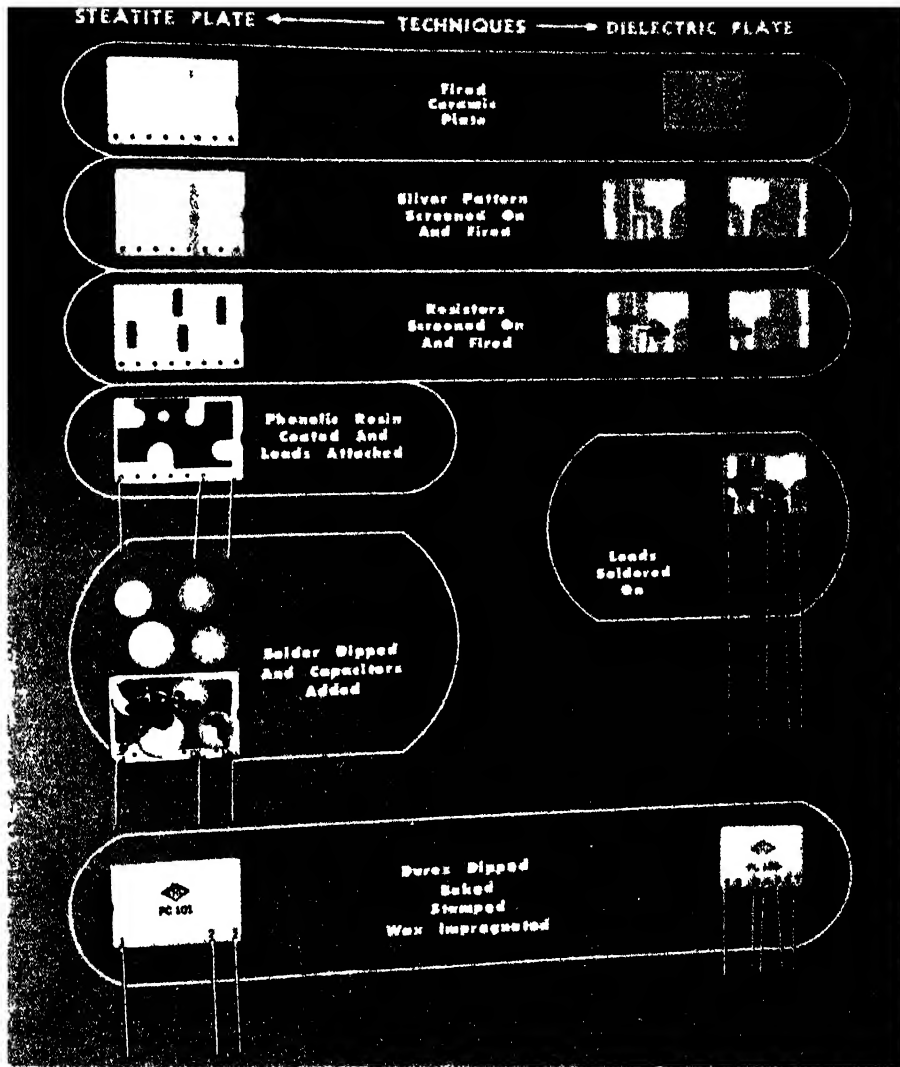


Fig 2 Printed circuit on ceramic-base materials. (a) Circuits in left column are screened on a steatite ceramic with barium titanate ceramic capacitors soldered to base. Circuits on right column screened on a high dielectric constant ceramic. The base plate is used

as the dielectric for the capacitor formed by screening a conductive area on opposite sides of the plate. (b) Schematic diagrams of circuits in (a). (Centralab Div., Globe Union, Inc.)

are used for deposition of materials to form resistances or capacitances. Photoelectrostatic techniques have also been developed, primarily for the fabrication of conductive elements.

**Screening method.** The screening method to form the circuit (conductors, resistors, capacitors, and in some cases low-value inductances) uses ceramic or glass as a base material (see Fig. 2). A silver paste, usually with a glass binder, is applied through a screen stencil to form the conductor pattern. The stencil is usually formed by a photographic process similar to that used for the photo-etching process previously mentioned. After the silver paste has dried, it is fired in a furnace at temperatures between 750 and 1350°F to bond the silver pattern to the base.

**Resistance components.** Resistive elements are usually formed by one of the three methods listed in the order of degree of precision required: (1) screened or sprayed, (2) applied as a tape, and (3) vacuum deposited.

Resistive inks are composed of various forms of carbon with a resin binder (such as phenolic or epoxy) and a solvent vehicle. This mixture is applied through a stencil to form a rectangular pattern between two conductors and then baked at a temperature between 150 and 600°F. The value of resistance is determined by varying the type of carbon used, the ratio of carbon to binder mix, the aspect ratio (the ratio of length to width) of the rectangular pattern, and to some extent, the thickness of the film. Resistive elements printed in this manner have wide tolerance limits, and it is usually necessary to adjust the resistive value by some means such as abrading the surface or one edge of the pattern. Thus the resistive element is formed with a value below that required and then adjusted to fall within the design tolerance limits. The stability of the resistance value is dependent primarily on the type of carbon used and the binder material. In general, resins cured at higher temperatures provide more stable resistance values. Thus an epoxy or silicone binder would provide a more stable resistance than that obtained with a phenolic binder. Printed resistors of the type mentioned above are used primarily in consumer products, such as radio or television sets, where the tolerance requirements and operating environment are not too severe.

In order to overcome the difficulty of producing resistive elements to close tolerances, and to increase the probability of producing a number of resistive elements on the same circuit base to reasonable tolerances, the tape resistor was developed. This form of resistor is capable of operating at temperatures up to 400°F. The resistor is fabricated by spraying metal-free asbestos paper with an ink composed of carbon, silicone resin, and solvent. The tape is applied to the circuit base and is then cured at 575°F for several hours to polymerize the resin. Curing of the resin binder also bonds the tape to the circuit base, thus forming the completed resistor. Resistance values between 10 ohms and 10 megohms  $\pm 6\%$  within an area of 0.13 by

0.30 in. are obtainable by changing the type of carbon used and the ratio of carbon to resin.

Fabrication of printed resistors by vacuum deposition is an expensive process, and there has been no general application of this technique except for precision resistive elements. Usually alloys with a low thermal coefficient of resistivity (such as nickel-chromium alloys) are deposited as a thin film on a glass or ceramic base by vacuum-evaporation techniques.

**Capacitance components.** Printed capacitors are fabricated as part of the conductor circuit pattern when a high dielectric constant ceramic, such as one of the titanates, is used as the circuit base material. Conductive patterns are screened on opposite sides of the circuit base to form the capacitor. The dielectric constant of the titanates varies widely with temperature; thus these capacitors are quite temperature-sensitive and are limited to circuits that accommodate wide circuit tolerance.

Most printed-circuit capacitors are disks fabricated from barium titanate with additives to obtain temperature stability. These ceramic disks are silvered on both sides in much the same manner as the circuit patterns were formed on a ceramic base. The capacitor is joined to the ceramic circuit base by the use of lead-tin solder containing approximately 3-5% silver to prevent the molten solder from dissolving the silvered conductors. •

**Plating process.** In the plating process (second only to etching in popularity for the fabrication of circuit conductors) a plastic laminate, such as paper-base phenolic, is first coated with a material that conducts electricity. This may be done by forming a 0.0001-in. silver coating on the surface of the laminate in much the same way that mirrors are silvered. The silver film is then coated with a



Fig. 3. Close-up of machine embossing a copper foil circuit pattern onto a phenolic-paper laminate. The copper foil, feeding into the machine from left to right, has a temperature-sensitive adhesive on the under surface. During the embossing operation a hot die cuts the circuit pattern, embeds the edges of the conductors into the phenolic, and temperature-cures the adhesive to form a bond with the base plate. (A. W. Franklin)

plating resist, usually an enamel, by a stenciling process, leaving exposed areas to form the circuit pattern. The plating process is similar to the plating of decorative metals. After sufficient thickness (usually 0.001–0.005 in.) of copper is deposited, the plating resist is removed by a solvent. The exposed silver film is removed by acid etching, leaving the much thicker copper plating to form the circuit pattern on the plastic sheet. For electronic circuits that may be subjected to high humidity and continuous application of voltages (such that conditions for silver migration exist), the chemical reduction of copper instead of silver is used to form the plating electrode. All the subsequent processes are as described previously.

**Mold and die processes.** Of the mold and die processes, embossing and stamping are the most popular (see Fig. 3). In general, these two processes are used for the fabrication of conductors or inductances. Usually copper foil is embossed on a phenolic laminate base plate, or powdered silver is stamped on a plastic sheet. Because of the relatively high cost of tooling and circuit fabrication limitations, these processes have not gained wide usage. [1 K.L.]

**Bibliography:** C. Brunetti (ed.), *New Advances in Printed Circuits*, Natl. Bur. Standards, Misc. Publ. 192, 1948; F. M. Hom, L. K. Lee, E. R. Gamson, and R. F. Newton, *Development and Application of Automatic Assembly Techniques for Miniaturized Electronic Equipment*, Wright Air Development Center Tech. Rept 55 230, 1955, *Proceedings of the Symposium on Printed Circuits*, Radio-Electronic-Television Manufacturers Assoc., 1955.

## Printing

For the Western World, printing had its beginning during the first half of the fifteenth century with the invention of printing from movable metal types. This invention is commonly accredited to Johannes Gutenberg whose name today is chiefly associated with the Gutenberg Bible, probably the most famous book ever printed. Printing of a crude sort, done from movable types, was produced in Holland some years before Gutenberg began his activities, however, and the Dutch ascribe the invention to one Laurens Janszoon Coster of Haarlem. The historical evidence, debated by generations of scholars, does not definitely establish any one man as the inventor. Fragments of this early printing, known as incunabula or cradle printing, have survived and are preserved in European and American museums.

Both Gutenberg and Coster were preceded by several centuries by a Chinese, Pi Shêng, who printed from movable types made of earthenware before the year A.D. 1100; however, Chinese use of the art then and subsequently had no perceptible impact on Europe.

Regardless of who the inventor may have been, the invention not only was the starting point for what today is a multibillion-dollar industry but was one of the most significant inventions in human

history. It was significant, even revolutionary, in two respects. With movable types as cast by Gutenberg (small rectangular pieces of metal of uniform height and body thickness, with a letter or other character in relief on one end), it became possible for the first time to assemble letters into words, words into lines, and lines into pages, and to ink and print the pages. Thereafter, when wiped clean of ink, the pages, lines, and words could be disassembled and the types returned to their respective compartments in the type case, ready to be used again in another job. Up to that time the western world had known only block printing—the cutting of crude pictures and lettering on a block or blocks of wood. The parts that were not to print were cut away so as to leave the lines of the picture and the lettering in relief. Such lettering, cut for specific wordage, was an integral part of the block and could not be disassembled and recomposed.

A second and more important aspect of the invention was that, again for the first time in western history, it made duplication by other than manual methods possible. Before the time of Gutenberg and Coster, the text of all books, as well as of smaller pieces, was lettered by hand. When, for example, a monastery desired a new or additional copy of the Bible, the scribes were set to work laboriously copying the Old and New Testaments, and possibly the Apocrypha as well, on parchment or vellum, with a quill pen or brush. Later the work would be rubricated and sometimes embellished with initials, ornaments, and even pictures, done in gold leaf and several colors. All this might require the collective efforts and talents of several successive generations of handicraftsmen, and in the end the result was a single copy. With the invention of printing, it became possible to produce identical copies, as many as the patience and energy of the printer and the extent of his technical facilities might permit; and this in turn made it possible to put more information into the hands of more people in less time and at lower cost, and thereby to spread literacy and learning more widely and rapidly, than had ever been possible before.

It was this unprecedented ability to duplicate, to produce many identical copies instead of a single copy, that caused printing in its early days to be associated with magic and known as the Black Art.

The present-day printing industry may be said to have reached its current status primarily as the result of four major developments, of which this invention of printing from movable types was the first. The second, some 400 years later, was the invention of the composing machine, which did mechanically and quickly what Gutenberg and his successors for generations had done laboriously by hand. Third was the application of power to the printing press, culminating in the development of the high-speed, web-fed, multicolor rotary press; and fourth, the application to printing processes of the camera—first to photoengraving, then to lithography, especially offset, and most recently, via phototypesetting, to the composition of type matter.

It has often been said that the key to the future in the graphic arts is the camera, and it is possible that this last development may in the end turn out to be the most important of the four. Meanwhile, as in every mechanized industry, other technical developments, some now in the experimental stage, some already in being, for instance, printing in which ink is "jumped" from type to paper across a gap, may revolutionize the industry, not necessarily overnight but on short notice, and thereby may render obsolete some or all of the procedures and techniques described in this article.

In the printing industry of today there are three major methods or processes in use for transferring ink to paper. These are (1) relief or letterpress printing, in which the printing surface, composed of type and printing plates, is in relief and the nonprinting parts are below the printing surface; (2) planographic printing, in which type matter and pictorial material, transferred (usually by photography) to a single plate, are printed from an even surface; and (3) intaglio printing, in which the parts that are to print, both typographic and pictorial, are cut or etched into the plate and are below the plate surface.

Letterpress, in addition to being used to designate relief printing, is also used in book publishing to refer to text matter or typographic content of books in contrast to illustrations.

To the above is sometimes added a fourth process, screen or stencil printing, in which ink is brushed or squeezed onto paper through a stencil that may carry either pictorial or typographic material or both. Although this process accounts for a comparatively small part of the total volume of printing, the introduction of mechanization in recent years has given it the largest percentage of growth of all the processes. There are also specialized processes of limited application, in copying machines, for instance, which are considered more fully in another article. See **PHOTOCOPYING PROCESSES**.

#### **RELIEF PRINTING (LETTERPRESS)**

Relief or letterpress printing is the oldest of the three major processes, with a history that dates back more than 500 years, twice that if its beginnings in China are included. Although other processes, notably offset, have shown a relatively greater percentage increase in volume in recent years and have taken over an increasing amount of work formerly done by letterpress, it still remains the basic process and the largest in volume of the three. It is used today for almost all newspapers, for books and magazines, and for the general run of advertising, commercial, form, and miscellaneous printing.

**Typesetting.** In the procedure usually followed in letterpress printing, copy, marked for size and face of type and for length of line, is sent to the composing room, where it is set in type, usually on one of the "hot metal" composing machines—Linotype, Intertype, or Monotype; or Ludlow Ty-

pograph in the larger type sizes. Foundrycast type for handsetting is also used but mostly for headings and the like. The resulting composition, including heads, body matter, captions, and incidentals, after having been proofread and corrected, is then assembled into pages, along with line engravings (printing plates made from pen-and-ink drawings and other art work done in line) and halftones (plates made from photographs, wash drawings and other art work done in continuous tone). See **COMPOSITION (TYPE)**; **PRINTING PLATE**.

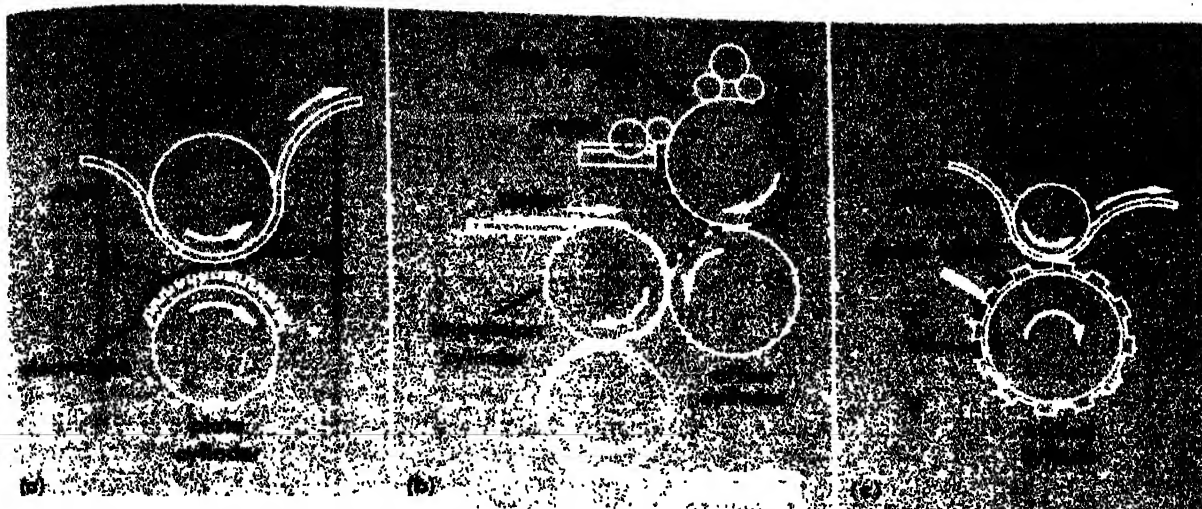
In the printing of large-edition newspapers, type and plates are assembled into pages and the pages each locked up in a heavy metal frame called a chase, which holds all elements of the page securely in place. A thick sheet of papier mâché, the stereotype matrix (mat), is then placed over the type page and squeezed down upon it by hydraulic or mechanical pressure, so that, when removed and dried, it constitutes a mold or matrix of the page. From this in turn a stereotype plate is cast, usually in semicylindrical form and about  $\frac{3}{8}$  in. in thickness. The curvature corresponds to that of the printing cylinder of the press on which it is to be used. If the page or part of the page is to print in a color or colors (in addition to black), a separate mat and plate are made for each color.

**Rotary presses (newspaper).** Because of the high production speeds required, the large metropolitan dailies are printed on web-fed rotary presses, which print from large heavy rolls of newspaper that pass through the press in continuous webs, and which carry the printing plates on the surface of printing cylinders. Each printing cylinder usually carries four or eight plates and thus will print four or eight pages on one side of the web at each revolution. An impression cylinder keeps the web pressed against the plates, which are inked by a set of rollers also tangent to the printing cylinder.

The large rotary presses consist of a number of printing units, each containing two printing cylinders (complete with impression cylinders) and each printing on a separate web, four or eight pages on one side and four or eight pages on the other. Any or all of these units may be brought into operation, depending upon the number of pages to be printed, and may be used for additional sections or for color printing. (The increasing use of color has been one of the outstanding developments in newspaper printing during the 1950s.) Eventually the webs from the several units are brought together and fed through a device, built into or attached to the press, which automatically gives them their down-the-middle fold, cuts them into individual papers with pages in consecutive order, folds the papers across the middle, and counts them—all this at speeds of 50,000 or more impressions per hour per press unit. Web speeds in the larger rotaries average about 1500 ft/min.

For dailies of smaller circulation which still must meet exacting delivery deadlines, a smaller type of rotary press is built, the tubular press—so called





Printing processes. (a) Relief. (b) Planographic. (c) Intaglio.

because the stereotype plates used are cast in cylindrical instead of semicylindrical form and are carried on printing cylinders of correspondingly smaller diameter. These presses consist of up to four units, each usually printing four pages, two on each side of the web. They can print 12- or 16-page papers at speeds of about 45,000 an hour.

For dailies of only a few thousand circulation and for weekly papers, there are several types of more or less specialized press, some printing from webs and some from sheets. Some of these operate on the flat-bed principle, and most print directly from the type pages without use of stereotypes or other form of duplicate plates.

**Rotary presses (magazine).** The big-edition, nationally circulated magazines, when done letterpress (*Life*, *The Saturday Evening Post*, and *Reader's Digest* are examples), are printed on web-fed rotaries generally similar in operating principles to those used for the large dailies, but with two points of difference: the presses make possible the printing of illustrations in full color from fine-screen halftones on both sides of the web, and a different form of duplicate plate, the electrotype (electro), is generally used instead of the stereotype. The electrotype is prepared by making a mold of the type page or halftone plate as the case may be, then suspending this mold in a bath of copper sulfate and sulfuric acid where, by electrolytic action, a thin shell of copper is deposited on it, and finally pouring molten type metal into this shell to back it up and give it the body and strength required to stand up on the press, afterwards planing it down to the required thickness. For electrotypes intended for very long runs, especially with colored inks (some of which react chemically on copper), a thin surface of nickel is electrodeposited first, followed by copper shell and back-up metal. The electrotype is more expensive and takes longer to make than the stereotype—from 2-8 hours—but is more durable and gives more faithful, finer-quality duplication of 120- and 133-line halftones in both black and white and

color such as are used in "slick paper" magazines, in much book work, and in the better grades of advertising printing.

Stereotypes may be made either curved or flat; electrotypes are made flat, and if for use on rotary presses, are curved by mechanical pressure.

Electrotypes and stereotypes have both long been used in book printing, particularly when the first edition is large and subsequent editions likely. A more recent development for book work is a duplicate plate made of rubber on a flexible mounting which can be bound around the printing cylinder of a rotary press. Duplicate plates of molded rubber, plastics, and even nylon are also in use for book, magazine, and commercial printing.

**Flat-bed presses.** Smaller-edition printing, when done letterpress (publications, booklets, pamphlets, business forms, folders, the general run of commercial printing), is produced mostly on one or another form of flat-bed press, so called because the type pages or duplicate plates (if such plates are used), instead of being locked on the surface of a printing cylinder, are positioned on a flat, sliding bed, horizontal on most presses, which moves back and forth under a cylinder. Most of this printing is done not on webs but on sheets of paper ranging in size from 12 by 18 to 44 by 64 in. and even larger. The cylinder is not a printing cylinder; instead, as it revolves, it seizes an edge of the sheet in its grippers and carries the sheet around with it, bringing it into contact with the form of type pages or plates as the bed slides through underneath, thereby printing one side of the sheet. Forms may consist of 4, 8, 12, 16, or more pages (they are generally in multiples of four) which are locked up together in a chase and printed in a single impression. As the bed of the press is returned to its original position, the cylinder is raised by cam action so that it will not come in contact with the form, while at the same time it delivers the printed sheet onto tapes or similar devices which carry it to the delivery end of the press and deposit it on the sheets previously printed.



Presses of this type are also known as cylinder presses and in most cases require two revolutions of the cylinder for each impression.

The sheets produced on flat-bed presses, after being printed on both sides, are folded by machine (or sometimes by hand) into signatures of 4, 8, 16, 32, or more pages, with pages in consecutive order, and the signatures subsequently trimmed to a predetermined page size. A printed piece such as a booklet or catalog may consist of one or several signatures, which are held together by wire staples, thread, flexible adhesives, or by mechanical binding devices made of metal or plastic materials.

Flat-bed presses are made in various sizes and models, including some that operate vertically. They do not have the speed of rotaries, production ranging approximately from 1500 or 2000 to 6000 impressions per hour, depending upon size of sheet and nature of work; but they are far more flexible in terms of short and long runs, different sizes of sheet, and ability to handle a wide range of work, and cost less to purchase and to operate. They are in general use in commercial printing plants and are capable of turning out a wide variety of printed pieces which vary in layout, structure, and quality in accordance with customers' specifications.

**Platen presses.** Small printed pieces such as envelopes, postcards, stationery, circulars, and small business forms are usually printed on platen presses—the familiar open and close or clamshell action type of press, common to printing plants generally, particularly the smaller ones, and generally known as job presses. They may be either hand or mechanically fed. Or, if required in large editions, these smaller pieces may be printed on large sheets from two, four, eight, or more identical type forms, plates, or sets of plates, the sheets being later cut up into the individual pieces. Often also, such jobs may be ganged up, that is, several different jobs run together in a single form and printed on the same sheet, to be subsequently cut apart.

**Special presses.** There are numerous variations on the types of press described above, including presses designed for specific repetitive jobs or for a specific type of work; but the three described, rotaries, flat-beds, and platens, are the types in widest use for general printing.

A specialized form of relief printing used for the printing of bags, food wrappers (including bread wrappers), acetate, cellophane, waxed papers, metal foil, and other materials having surfaces not easily handled by ordinary letterpress methods is aniline printing, now more generally known as flexography. The earlier name came from the fact that aniline dyes were (and are) used for pigments in the inks. These inks are liquid, dry quickly, and because of their use on paper and other substances that are directly in contact with foods, must be noncontaminating. Printing is done from flexible rubber plates carried on the printing cylinders of web-fed rotaries. See **PRINTING PRESS**.

## **PLANOGRAPHIC PRINTING**

Planographic printing is printing from an even surface—one in which the printing portions are neither above the nonprinting portions, as in letterpress, nor below the nonprinting portions, as in intaglio.

A second distinguishing feature is that the process (lithography), by which all but a very small percentage of planographic printing is produced, is based on the simple fact that grease and water will not mix.

Lithography, the invention, possibly by accident, of Alois Senefelder, dates only from 1798, nearly 350 years after Gutenberg. The term means literally stone writing or stone printing, and for the first 100 years of its existence lithography was just that.

**Direct lithography.** In the original process the copy to be printed (lettering, music, or art work) was either drawn by hand in reverse (right to left) on the surface of a slab of porous stone with a grease crayon or in greasy ink, or was transferred to the stone by rubbing, having first been drawn with a grease crayon on transfer paper having a special surface. The surface of the stone was then sponged with a solution of gum arabic in water to render the nonprinting portions receptive to moisture but repellent to greasy ink and the printing portions receptive to grease but repellent to moisture. The surface, after being dampened with water, was then rolled with the greasy ink, which adhered only to the printing image; paper was laid over it, and a print made by pressure.

This process, known as direct lithography, is little used commercially today but survives as a fine arts process for the making of lithographs, prints made from drawings or lettering done manually on the stone and printed on a hand press.

**Indirect lithography.** The lithographic process, however, not only continues but flourishes in the form of indirect lithography—offset lithography (offset for short). The clumsy stones have been replaced by thin, flexible plates usually made of zinc or aluminum but sometimes of plastic or plastic-coated paper. The plates usually have a fine-grained surface, though nongrained plates are also in use; the slow, stop-and-start flat-bed presses on which the stones had to be printed have been superseded by presses having a continuous rotary motion; and photographic methods have almost entirely supplanted the old manual and hand-transfer methods of putting the image on the plate.

Although offset has not as yet caught up with letterpress and rotogravure in the printing of metropolitan newspapers or nationally circulated magazines (experimental work toward that end has been in progress for several years and is still in progress), it is being used for almost the entire range of commercial and advertising printing, including window and counter displays, posters, road maps, packages and containers, labels, and picture postcards. It is also being increasingly employed for books, book jackets, and for publica-

tions of small or moderate circulation; and in conjunction with "cold type" (type produced on machines operating on the typewriter principle as opposed to "hot metal" type), it has taken over a vast mass of printing in which budgetary considerations are paramount. In the newspaper field, as yet, its use has been confined to a comparatively small number of weeklies and a few smaller-city dailies.

**Offset presses.** In construction and operating principle, offset presses differ widely from their letterpress counterparts. The typical press contains four cylinders of uniform diameter. This diameter and the length of the cylinders vary with the size of plate to be handled. First of the four is the printing cylinder, which carries the flexible metal printing plate clamped tightly around it. Bearing on this, one on each side of the cylinder, are two sets of rollers: the water rollers, which moisten the face of the plate as the cylinder revolves, and the ink rollers, which supply the ink. Here again the grease-and-water principle applies; the water is retained on the nonprinting parts of the plate but runs off the greasy printing image, while the image holds the ink, which is also greasy, but repels the water.

The printing cylinder, notwithstanding its name, does not print the image on paper; in fact, it never comes in contact with the paper. Instead, it prints the image on the surface of the second cylinder, an intermediate cylinder, known as the rubber blanket cylinder because its surface is rubber. This cylinder is generally located below the printing cylinder and is tangent to it.

To one side (usually) of the rubber blanket cylinder is the third cylinder, the impression cylinder, the purpose of which, as in a letterpress rotary, is to hold the paper against the rubber blanket cylinder as the cylinders revolve. The paper, in the form of either sheets or web (there are both sheet-fed and web-fed offset presses), feeds in between the impression cylinder and the rubber blanket cylinder, and printing takes place at the point at which the two are in contact. Thus, the printing image is transferred to, or offset on, the paper, not printed directly on it by the printing plate; hence the terms *offset* and *indirect lithography*. The fourth cylinder, which is usually below the impression cylinder and revolves against it, is the delivery cylinder and carries the printed sheet or web to the delivery end of the press.

Like the larger letterpress rotaries, an offset press may be made up of several printing units. Thus, a press designed for four-color process printing may have four units in tandem, one printing the yellow, one the red, one the blue, and one the black; the colors being printed successively (although not necessarily in that order) on the sheet or web as it travels through the press. Smaller presses are made in single-color and two-color models.

Attempts have been made from time to time to eliminate the intermediate (rubber blanket) cylinder in favor of printing direct from plate to paper.

As yet, however, it continues to hold its place, mainly for the reason that, having an elastic instead of a rigid metal surface, it can print halftones, even fine-screen halftones, on rough-surfaced papers, on metal (as in the printing of tin), and on other uneven surfaces, including canvas, wood, and leather. This is either mechanically impossible or commercially impractical with letterpress. It also contributes to, or at least does not interfere with, the continuous rotary motion of the press, and with albumen plates particularly, may contribute to the life of the plate by enabling the photographically imprinted image to stand up longer on the press than it would if in direct contact with the paper.

**Dry offset process.** Dry offset is a process which employs the offset principle but in which the planographic printing plate is replaced with a plate having the image in relief, thereby eliminating the need for the water rollers, which on occasion have been the source of production difficulties. Although this process is not generally available, it holds definite possibilities for the future.

**Planographic composition.** Offset production differs from letterpress production in several respects, beginning with composition. If type produced on "hot metal" machines is to be used, the type is set, proofed, read, and corrected as in letterpress, after which repro proofs (proofs for photographic reproduction) are taken. These proofs are usually pulled on a coated or dull coated paper on a special type of proof press, and are as nearly perfect mechanically as it is possible to make them. If "cold type" composition is to be used, the copy is typed on smooth-surfaced white paper, either on a typewriter or on one of the several "cold type" composing machines, and the typed sheets are used in the same manner as repro proofs.

In the next stage the repro proofs or typed sheets are cut up and pasted down together, usually with rubber cement, on a carefully laid out sheet, with body matter, heads, and captions in the exact position they are to occupy in the printed product, the result being what is known as a *keyline layout*, or more often in the trade, as a *mechanical*. Such mechanicals may consist of a single page, a spread (two facing pages), or of 4, 8, 12, 16, or more pages, arranged as in a form in letterpress. Line illustrations, if drawn to the proper scale, may be mounted in position on the mechanical; if too large for this, they may be photographed down (reduced) to the proper size and the photographic prints mounted in place instead, or separate line negatives may be made. Spaces to be occupied by halftones are either ruled off on the mechanical or are indicated with rectangles of black paper cut to the exact size the halftones are to be and mounted in their several positions.

Headline and text material may be composed for offset production on phototypesetting or photocomposing machines, thus eliminating the necessity for making and photographing repro proofs or typed sheets. However, if photographs, wash drawings, or

other kinds of tonal art work are to be used as illustrations, they are photographed separately through a halftone screen, and halftone negatives are made.

All negatives, headline and text, halftone and line, are carefully inspected on a light table, and "pinholes" or other imperfections are opaqued out, that is, painted out on the negative with an opaque purple or black paint.

**Planographic plates.** In letterpress, separate plates are made from the halftone negatives. In offset, the halftone negatives and the line negatives of type and line art are assembled and taped down together with transparent gummed paper on a "flat," the negatives being positioned exactly as the type and pictorial matter are to appear in the finished product. The flat is a sheet of opaque paper, usually colored, on which a careful layout corresponding to that of the mechanical has been made and openings cut for the line and halftone negatives that are gummed down on it so that the paper partly masks out the nonprinting portions. In other words, the flat is a composite negative made up of the negatives of all the material, both typographic and pictorial, that is to appear in the page, spread, or form, as the case may be. In color work a flat is made for each color, the flats being carefully registered one with another.

In all negatives, whether line or halftone, the black and white values are reversed. Type proved in black ink on white paper, for instance, will appear white and transparent against an opaque black background in a negative.

**Preparation of plates.** The flat is next laid over the flexible metal plate, the surface of which has previously been coated with a light-sensitive emulsion (presensitized plates are widely used in offset plants) and the two are locked together in a vacuum printing frame. These vacuum printing frames come in several forms, a common one being similar to a large table with a steel-framed glass top hinged on one side; this top is raised, the flat and plate inserted, the top locked down, and the air exhausted to ensure perfect contact between flat and plate. The image is then "burned in" with powerful lights, which shine down through the negatives to the plate, hardening the emulsion under the transparent lines and dots and rendering it insoluble in water. Nonprinting areas, which are shielded by the opaque parts of the negatives, are not affected by the action of the light and remain soluble.

After exposure, which varies in length of time according to the nature of the work to be printed, the flat and plate are removed from the printing frame and separated, the face of the plate is coated with a special ink, and the plate is then washed. The ink adheres to and brings out the printing image but washes away from the nonprinting areas, carrying the emulsion with it and leaving the metal exposed. The printing image is then fixed, and subject to final change or correction, the plate is ready for the press.

When the flat has been completed but before it is put into the vacuum printing frame, a photo-

graphic print is made from it, usually on blueprint paper or its equivalent; inexpensive photographic papers are used which can be developed in water. This print, folded down into a signature with pages in consecutive order, is the equivalent of a final proof on which the author, editor, or production manager may, within limits, indicate his last corrections or mark his approval.

**Deep-etch plates.** The procedure described above applies to the making of albumen plates, in which egg albumen is the vehicle that carries the light-sensitive salts, and which are widely used. For sharper contrast and greater detail in halftones, however, and particularly in the more exacting forms of color work, the offset printer may make use of so-called deep-etch plates. For these, photographic positives instead of negatives are used in the flats, and the printing image is etched slightly below the surface of the plate, thereby enabling the plate to carry more ink than is mechanically possible with albumen plates. Deep-etch plates are also more durable than albumen plates. The term deep in this connection is a misnomer; actually the depth of etching is so slight that it can barely be felt with the fingernail, and the plate is classified as planographic, although in fact it is slightly intaglio.

A special form of deep-etch plate, the bimetal plate, has come into use in recent years. These plates are composed of two metals, one of which has a special affinity for water and the other for ink, the former constituting the surface of the plate. One combination used is a wash of chrome over copper. In this the image is etched through the chrome to the copper, thus taking advantage of the special qualities of the two metals. There is also a trimetal plate of chrome over copper backed up with steel or zinc.

For use in the flats for deep-etched plates, repro proofs are sometimes pulled on thin sheets of acetate on a special press which prints the image in exact register on both sides of the acetate to ensure opacity, and these acetate sheets are used in the same way as photographic positives. Similarly, the negatives or positives of type matter produced on phototypesetting or photocomposing machines may be used, instead of negatives of repro proofs, in the making of flats or mechanicals.

**Place of offset lithography.** It has been mentioned that in letterpress, jobs running to very large editions may be printed from two, four, eight, or more identical plates or sets of plates, usually electrotypes, with a corresponding reduction in press running time required. In offset, the equivalent of this duplicate plate technique is obtained by means of the step-and-repeat machine, whereby identical images, precision-spaced at fixed distances apart (both sideways and up and down), may be projected photographically onto the surface of a single large plate. With this device, sets of large color plates carrying multiple images may be made which will register exactly, plate to plate and image to image, when the colors are printed successively, one over the other.

Of the three major printing processes—letterpress, offset, and gravure—offset is by a considerable margin the newest, having come into general use in the United States in 1912. Its invention, or at least its application to the printing of paper (the offset principle had been employed earlier for lithographic printing on tin), is jointly credited to Ira W. Rubel and to the Harris brothers, A. F. and Charles. Rubel and the Harrises, working separately and independently, appear to have arrived at approximately the same results at approximately the same time. Like Senefelder's invention of lithography, the invention of offset may have been at least partly accidental.

Although the total volume of work produced by lithography is still materially below that produced by letterpress, its percentage increase has been very much higher, 105% as against 38% from 1947 to 1954. The increase in volume over the same period totaled \$665,000,000 for letterpress and \$550,000,000 for lithography. In gravure, which runs third in total volume, the percentage increase from 1947 to 1954 was 205%; the volume increase was \$114,000,000.

#### PHOTOGELATIN PRINTING (COLLOTYPE)

A process which belongs in the planographic category, although not literally planographic, is photogelatin printing or collogtype, a distinguishing characteristic of which is that it uses no screen. In its modern version—it originally was mainly a hand process in which production might amount to no more than 100 impressions per day—the plate, usually of aluminum, is coated with a layer of light-sensitive (bichromated) gelatin, over which is laid an unscreened photographic negative of the copy to be reproduced. When the two are exposed to light, the gelatin is affected in varying degree in proportion to the amount of light admitted to it through the dark, medium, and light portions of the negative. The plate is given further treatments, including immersion in a glycerin solution, and the end result is a printing image which varies not only in thickness (in terms of thousandths of 1 in.) but in moisture content and consequent ability to repel ink or hold it in varying amounts. The plate is kept moist and the pressroom maintained at the proper degree of humidity by controlled humidification.

Photogelatin or collogtype is essentially a short-run process; runs of only a few hundred prints are common, with 5000 to perhaps 10,000 as a practical maximum, but is used for a fairly wide range of advertising work in both black and white and color, including counter displays, posters, blow-ups (enlargements) of advertisements and pictorial copy, and facsimile reproductions of photographs. Because of its screenlessness it has been used to great advantage in reproducing ancient manuscripts which either were falling to pieces from age or were likely to suffer from handling and exposure to the air. Of all processes, it gives, or is capable of giving, the most faithful facsimile reproductions.

Printing, formerly done on stop-cylinder presses from flat plates of heavy glass, is now done on

sheet-fed rotaries. Production speeds vary according to the nature and size of the copy and the surface on which it is to be printed.

#### INTAGLIO PRINTING

The parental process in intaglio printing (professionally pronounced in-taggle-io) is steel and copperplate engraving, which rivals letterpress in age, going back in Europe at least as far as the time of Gutenberg and probably farther, and, in China, possibly to the time of Pi Shêng. According to D. B. Updike the first book decorated with copperplates was printed as early as 1477 (Gutenberg died in 1468) and books with copperplate pictures and title pages were fairly abundant in the late sixteenth and the seventeenth centuries.

**Engraving.** In this process the artist took a thin but rigid plate of copper or steel, drew or sketched his picture on it, usually in right-to-left reverse, and then cut the lines of the picture, and of the lettering if any, into the face of the metal with a small V-pointed graver or burin. After the cutting was completed, the plate was inked over its entire face and then wiped, so that the ink was removed from the surface but was retained in the incised or engraved lines. This is the original and literal meaning of the term engraving, which today is applied generally to relief plates in both line and halftone. Paper is then laid over the plate, paper and plate are inserted in a crude form of hand press, and the ink is transferred from engraved lines to paper by a combination of pressure and suction.

Although it is extensively imitated by a form of raised printing, the process survives today in the making of plates for formal calling cards, engraved stationery, formal announcements such as wedding invitations, certain forms of certificates, and in artists' etchings. In these last, the surface of the plate is usually coated with wax and the drawing done in the wax with a sharp-pointed instrument which penetrates through to the copper. Thereafter the plate is etched with acid, the lines being "bitten down" into the metal; the wax, which protects the nonprinting parts from the acid, is removed, and the plate is inked, wiped, and printed.

Intaglio printing from steel plates is used in the U.S. Bureau of Engraving and Printing for the printing of stamps, paper money, and bonds of large denominations. Stock certificates and bonds of commercial organizations are usually printed by private firms specializing in steel-plate reproduction. Serial numbers on money, stock certificates, bonds, and other documents are imprinted by special letterpress attachments.

**Gravure processes.** Out of this basic traditional technique have grown the three processes which today are known collectively as gravure—photogravure, rotogravure, and sheet-fed gravure. Of these, the largest and the most important commercially is rotogravure, which is done from large, engraved, copper-surfaced or copper-jacketed cylinders carried on web-fed presses of the rotary type, and is used for newspaper roto sections, for some

of the big-edition magazines, and for big-edition commercial work. (*This Week* magazine, *The American Weekly*, the Sunday magazine and book review sections of *The New York Times*, and certain pictorial sections of the catalogs issued by the large mail-order houses are examples.)

Because of the time required for plate preparation and the relatively high initial costs, rotogravure is essentially a large-scale production process, best suited to work that runs into hundreds of thousands or millions of copies. Photogravure is virtually a hand process, best suited to limited or deluxe editions in which reproductions of the finest quality are called for, and is capable of beautiful effects. Sheet-fed gravure, which has been more highly developed and is more extensively used in England and on the Continent than here, lies in between; it is suited to runs of moderate or medium length, and its costs are not so high as to put it out of competition with fine-quality letterpress work and with deep-etch offset.

In reproducing photographs and tonal art work, the gravure processes are capable of a depth and quality of tone superior to that of any other process, but type reproduction suffers because all matter, typographic or pictorial, going on a rotogravure cylinder or a sheet-fed gravure plate must be screened, and readability of type may thereby be affected. For this reason type matter printed letterpress and gravure-printed illustrations are sometimes used together in the same publication or advertising piece.

**Printing surfaces.** Without going into the technical details of screen and carbon tissue, it may be said that the printing surface of a gravure cylinder or plate as used for most monochrome work (black and white, or as often, sepia and white) differs from the surface of a letterpress halftone in three particulars: (1) the dots composing the printing image are below the surface instead of in relief; (2) instead of varying in size or area as in a halftone, the dots are all of the same area; and (3) although the dots are of uniform size surface-wise, they vary in depth, this depth being greater or less according as the parts of the copy printed from them are respectively dark or light. For dark parts (blacks or dark grays), in other words, the tiny, almost microscopic cups or wells carrying the ink will be deeper, will hold more ink, and consequently will apply more ink to the paper than will the shallower cups or wells which print the white or light gray portions. (Lest the terms cups and wells prove misleading, it should be said that with the 150-line crossline screen used in most monochrome gravure work, there are some 22,500 of them per square inch, and that the depth of the deepest will not exceed 0.002 or 0.003 in.)

It is this ability of the gravure process to print from what is, in effect, an ink film of varying thickness instead of from one of uniform thickness, plus the ability of the deeper dots to apply more ink to the paper than is possible with other processes, that accounts for the range of tonal values and the depth of tone characteristic of gravure-printed

reproductions. Because of the amount of ink applied, the dots on the darker portions frequently spread and overlap one another on the paper, obscuring the screen structure entirely and giving the effect of nonscreened continuous tone.

For color work in gravure, a process of comparatively recent development is used, the Dultgen or News-Dultgen Halftone Intaglio Process, in which, by the use of a reverse halftone screen and two photographic positives, dots are produced which vary in size as well as in depth. This combination of different-sized dots, as in relief halftones, with different-depth dots, as in crossline-screen gravure, while suited to monochrome work, is used mostly for three- and four-color process printing, where it gives added detail, fidelity, and color range to the reproduction.

**Printing operations.** In the printing operation, the gravure cylinder or plate either rotates through a trough of almost liquid ink, entirely unlike the thick, tacky letterpress and offset inks, or has the ink sprayed upon it, the ink being held on the surface as well as in the etched cups. As the cylinder continues its rotation, it passes under a thin, flexible steel blade or scraper, the doctor blade, which extends the entire length of the cylinder, bearing at an angle against it, and wipes or scrapes the ink from the surface, leaving it only in the etched cups. The blade thus does mechanically what the operator does manually in copperplate engraving. To minimize wear and the possible effect of small nicks, the blade is made to oscillate lengthwise against the cylinder. After being wiped, the cylinder rotates further and comes in contact with the paper, which is held against it by an impression cylinder, and printing takes place at that point.

Like the larger letterpress rotaries and offset presses, a gravure press may be made up of several units which may be used either for the printing of additional pages or for color printing. Rotogravure press speeds for monochrome work run around 20,000 or more impressions per hour.

Aside from sheet- as opposed to web-feeding, sheet-fed gravure differs from rotogravure mainly in the fact that the printing image is usually etched into a thin, flexible copper plate which is clamped around a printing cylinder in much the same manner as an offset plate. As compared with rotogravure, it is a quality as distinguished from a large-scale production process, is much slower (up to 7000 impressions per hour), will print on a wider range of stocks, and rightly handled, will produce results than can hardly be surpassed by any other process.

Sheet-fed gravure is sometimes referred to as photogravure, but this term is more generally and more correctly applied to the intaglio process mentioned earlier in which the plates are made and the printing done largely by manual methods. Photogravure antedated both roto and sheet-fed gravure. In photogravure, instead of a screen being used in plate making, the plate is lightly covered with fine powderlike particles of asphaltum.



an acid resist, which are floated on it by air in a special container. The plate is then heated to bake in these particles on the surface of the metal, the particles forming the equivalent of an acid-resistant screen or grain pattern; carbon tissue carrying a photographic positive is placed in position over it, and the plate is etched in much the same way as a plate for sheet-fed gravure. Production, done on a hand press, runs to 100 or fewer sheets per day.

### SCREEN (STENCIL) PRINTING

Of the stencil printing processes referred to earlier, there are two that should be mentioned here: mimeographing and silk-screen printing. The former is actually a copying or duplicating rather than a printing process, although capable of producing up to perhaps 5000 copies, and is widely known through its use in business offices and military circles. The stencil is usually typed on a typewriter. Lettering, diagrams, and drawings may be added by drawing or tracing them on the stencil with a stylus. Many small publications, such as church and company newsletters, and an enormous variety of releases, publicity handouts, and miscellaneous short-run copy of all types are handled by this technique. Both stencil cutting and printing can be done by office help, making it one of the most economical of processes.

More important from a printing standpoint is the silk-screen process, which is used for a wide variety of advertising, including posters, window and counter displays, and the like, and for printing on wood, metal, cloth, plastic, and other nonpaper surfaces. It also finds use as a fine arts print process and has many enthusiastic do-it-yourself practitioners. Essentially, the process consists of a screen, usually of 51-gage silk, on which, by any of several techniques, a design is traced or lettered and the nonprinting parts painted over with an ink- and water-resistant lacquer. This stencil is carried, tightly stretched, in a frame; the ink, which actually is a paint rather than an ink, is carried in the same frame; and by means of a squeegee the ink is pulled back and forth in the frame, printing at each pull an impression on the paper or other surface, which is held in position on a table underneath the frame. Printing is usually done in flat, opaque colors, one color at a time. Production by manual methods is slow, only a few hundred impressions per hour, but with power-driven presses, which are being used in constantly increasing numbers, a rate of production of 1200-3500 impressions per hour can be achieved. The versatility of this process makes it suited to an almost limitless number of applications; consequently, it is the most dynamic branch of the graphic arts industries. [L.B.S.]

**Bibliography:** American Photoengravers Association, *The Art of Photoengraving*, 1952; T. F. Carter, *The Invention of Printing in China and Its Spread Westward*, 2d ed., 1955; B. Dalgin, *Advertising Production: A Manual on the Mechanics of Newspaper Printing*, 1946; T. Dreier, *The Power of Print and Men*, 1936; L. Flader and

J. S. Mertle, *Modern Photoengraving*, 1948; O. W. Fuhrmann, *Gutenberg and the Strasbourg Documents of 1439*, 1940; L. C. Gandy, *The story of lithography*, *Lithographers' Journal*, suppl., 1940; *Graphic Arts Production Yearbook*, 1959; C. L. Helbert (ed.), *Printing Progress: A Mid-Century Report*, 1959; D. Hymes, *Production in Advertising and the Graphic Arts*, 1958; D. C. McMurtrie, *The Book*, 1937; D. Melcher and N. Larick, *Printing and Production Handbook*, 2d ed., 1956; G. J. Mills, *Sources of Information in the American Graphic Arts*, 1951; J. C. Oswald, *A History of Printing; Its Development through Five Hundred Years*, 1928; *The Penrose Annual*, 1959; T. B. Stanley, *The Technique of Advertising Production*, 2d ed., 1954; V. Strauss (ed.), *The Lithographers Manual*, 1958; A. A. Sutton, *Design and Makeup of the Newspaper*, 1948; D. B. Updike, *Printing Types: Their History, Forms, and Use*, 2 vols., 2d ed., 1937.

### Printing in color

The art and craft of embellishing designs, pictures, and typographic pages for a more pleasing effect than obtained in black-and-white; in addition, for pictures which are more nearly representative of the original object or painting. Long before printed pages supplanted hand-written manuscripts, scribes ornamented or "illuminated" their pages with elaborate decorations (or rubrics, from the red colors often applied to titles and initial letters). Early printing also had illustrations tinted in color by hand. Wallpaper painted in colors dates back to 200 B.C., when the Chinese decorated rice paper with scenes from nature.

After the development of lithography by A. Senefelder about 1798, artists found it relatively easy to produce colored copies by working directly on the lithographic stone, using a number of different colors to achieve their results. A painter and etcher named J. C. Le Blon first conceived the possibility of producing a wide range of color effects by using only three basic colors. He attempted to commercialize his efforts in 1704, but three-color printing in the modern sense was developed after color photography. In the late nineteenth century, following the invention of the halftone screen, colored halftones were developed. The first three-color process illustrations similar to present-day methods are generally credited to W. Kurtz in 1893, although a number of workers in the field, S. H. Horgan, F. E. Ives, and L. E. and M. Levy, made significant contributions.

Three- and four-color process printing is closely allied with color photography; the fundamentals of color reproduction are much the same for all the printing processes. Just as a halftone printed in black on white paper is an optical illusion giving the observer the impression that he is seeing the various gradations of tone in the original photograph, a colored halftone gives the illusion that a wide range of colors is present. Actually, only three colors are used (black is added in the more commonly used four-color process). The halftone



dots in each color are printed at different angles, so that they fall one alongside another and overlap to form combinations of many colors. Inks for process printing are transparent tones of red, blue, and yellow. See **INK**; **PHOTOGRAPHY, COLOR**.

**Separation negatives.** Since three plates (four for four-color process) must be obtained in order to print the proportionate parts of the different colors, first steps involve breaking down a colored original into three (or four) separate photographic images. These are termed separation negatives. In the process, the original colored object, painting, photograph, or transparency, is positioned before the lens of a large copying camera. Over the lens is placed an orange-red filter which allows light rays of that color only to pass; thus, the red portions of the original are represented by tones of gray in the negative. In like manner, another piece of film is placed in the camera and another exposure made with a green filter over the lens. This negative now contains the green portion of the original. Again, a third piece of film is exposed through a deep-blue filter, this negative giving a record of the blue areas in the original. (For four-color process, a fourth separation must be made; see below.) Standard process filters used might be Wratten filters A25 (red), B58 (green), and C5-47 (blue).

Because the light-transmitting ability of these filters, as well as the sensitivity of the film, varies with the color, exposures must be regulated to obtain a set of separations which are properly balanced with each other. As an aid to the operator for correctly judging the results, a gray scale is photographed along with the original. This is a strip of paper or film with approximately ten steps in neutral shades of gray from black to white. In properly exposed and developed color separation negatives, the gray tones will match in all three. Typical exposures might be, for a red filter negative, 20 sec; green filter negative, 24 sec; and blue filter negative, 6 sec. Actual exposures would depend upon the lighting, camera settings, and kind of film used. A transmission densitometer may be used for more exact checking of results.

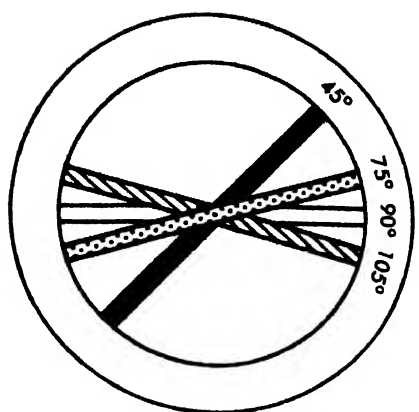
**Black printer.** Although it is theoretically possible to print the full range of tones using only the three process colors, most operators include a black printer to add detail and contrast to the printed reproduction. To make the separation negative for the black printer, the same general procedure used for the other separation negatives is followed. As a filter for this separation, the best choice (called split filter) is to use all three of the previous filters, one at a time, with exposures for each running from 50 to 100% of that used for each filter on the individual separations. Experience and judgment will determine the exact time. The object is to eliminate all but the major dark lines and shadows in the finished plate. A heavy black printing plate would interfere with clean, clear printing of the other colors.

Some operators prefer to use a single filter, such as a Wratten No. 8 (yellow), or other choice, depending upon the nature of the subject. A black printer separation negative made with a single filter usually requires more handwork for a satisfactory result. A third method of shooting the separation for the black printer uses infrared film and a filter (Wratten No. 88A) which transmits infrared rays. This method will not give best results with all subjects, and is usually restricted to paintings and pastel drawings. Perhaps the first choice for most purposes is the split-filter method mentioned.

**Screened positives.** From each of the separation negatives, a positive print must be made. A positive represents tonal values reversed from those in the negative, so the positive made from the red-filter negative will represent all colors except red, in other words, minus red. White light minus red leaves blue-green, the color in which this positive should be printed; in process work this color is called cyan. The green-filter negative produces a positive which must be reproduced in minus green, which is bluish-red or magenta. The blue-filter negative in turn gives a positive which must be printed as minus blue, or yellow. When the three positives are brought together one over the other in exact alignment, the original subject is re-created.

The steps followed from separation negatives to the combined positive prints vary according to the particular graphic arts process used. In the case of photographic prints, these positives are prepared on transparent films which are dyed in the respective process colors and the images superimposed to give a color print. For offset lithography, the separation negatives are each photographed through a halftone screen to give a screened positive from which three deep-etch plates are made. These plates are printed one at a time, superimposed in register, in the proper color for the final reproduction. In photoengraving, it would be necessary to make continuous tone positives from the separations, by contact printing, then make halftone negatives in the camera, after which the photoengraving plates would be made. It is also possible in photoengraving to make the separation negatives with the halftone screen in the camera, thus obtaining screened separation negatives in the first step. Exposures run considerably longer with both filter and screen between the lighted original and the film. This is known as the direct method, but does not permit as much control as the indirect method outlined above. For gravure, succeeding operations vary according to the exact system used; in conventional methods a special screen is used at the time cylinders are prepared. See **PRINTING**; **PRINTING PLATE**.

**Screen angle.** As each of the screened shots is made, it is necessary to change the angle of the screen ruling. This will allow an optical blending of the different colored dots to give the effect of the colors in the original; it also prevents a moiré pattern which appears when one screened image is



| yellow       red  
 blue       black

Screen angles used in four-color printing.

printed over another. Usually, screen angles are selected to put the strong color at an angle least noticeable to the eye, whereas the weaker color is used at the angle most pronounced to the eye. Thus, the black printing plate will be made with the dots running at an angle of 45° (as in the normal black-and-white halftone), and the yellow printing plate with the dots vertically and horizontally, or 90°. Magenta and cyan plates will have the dot angles between the other two, at 75 and 105° (see illustration).

Where only three colors are used, the cyan plate will probably be at 45°, the magenta at 75°, and the yellow at 105°. Some operators may also prefer to make the yellow plate with a different halftone ruling to further lessen any problem with moiré pattern. For example, the yellow plate may be made with a 150-line screen, and the other plates with the normal 133-line.

**Angled prints.** A simplified method of screening separations (as for coarse-screened newspaper work) calls for carefully balanced positive prints to be made on photographic paper of each of the separation negatives. These black-and-white photographs are turned to the proper angle before the camera, and halftone negatives are made for each. If the prints are relatively small, they can all be mounted together in front of the camera, each at its respective angle, and one shot made of the group. The developed negatives are cut apart and printing plates made in the usual fashion.

**Deficiencies of pigments.** In spite of the high level of fidelity to the original obtainable in the color separations using filters as described, no pigments, dyes, or printing inks can produce subtractively to the same degree that filters can additively. The ink for a given plate should be the color complementary to the filter used for the corresponding negative. Each ink should absorb only one-third of the color spectrum. Unfortunately, inks cannot presently be manufactured which will give ideal results in the printed image. Each ink tends to lap over and muddy up the other colors. To com-

pensate for this deficiency, it has always been necessary to distort the separation negatives by a certain amount of handwork, such as darkening some areas and lightening others in order to come closer to a facsimile reproduction of the original with the inks at hand. See COLOR; PIGMENT.

**Color-correction masks.** A new technique called masking has gradually replaced handwork to the point that most color work today is accomplished in this manner. A mask is a photographic image superimposed over another photographic image to alter its transmission characteristics. Masks may be used to change the contrast or to change the color balance of the original. An operator may choose from a number of masking methods, which may involve one or more masks, according to the final effect desired. Eastman Kodak Company and the Lithographic Technical Foundation have done a great deal of work on the use of masks in color reproduction.

**Added-color plates.** Occasionally it becomes necessary to use one or more extra colors in addition to the standard four-color process inks. This may be caused by a client's insistence upon a particular color match with that of his product, or in cases where metallic effects are wanted; for example, bronze, gold, and silver. Flat color backgrounds and borders are better handled by a separate printing, rather than by attempting to get the desired effect with process printing. Lithographers formerly used extra plates as a standard practice to enhance the appearance of the final reproduction. Light tints of pink, blue, and gray were commonly used. With improved techniques, these extra plates are not a standard part of process work, but additional flat colors are still used by all the processes.

**Electronic scanning.** In 1950, the Time-Life laboratories working with Eastman Kodak Company introduced an electronic device for scanning color transparencies and automatically producing a set of four-color separation negatives. Corrections are incorporated into the separations by the equivalent of electronic masking. Similar mechanisms have been announced by other groups, including the Acme Color Separator, the RCA-Interchemical Color Corrector, and the Hunter-Penrose Autoscan. These machines work on different principles, but all are used in an attempt to speed up and improve former means of getting color-balanced separations for process work. The Time-Life laboratories have set up scanning units in a few cities to provide, on a commercial basis, electronically produced sets of color-corrected separations for lithographers and photoengravers. No doubt, future developments will make such methods common; however, most color work is being carried out at present as outlined in this article.

**Short-run three-color system.** Within the last few years Eastman has developed a simplified method designed to produce color reproduction between the photographic color prints and the standard procedures outlined above. Standardization of

inks, simplification of routine, automation of operations, and careful control of all steps are characteristics of the process. This Kodak system does not supplant the normal systems, but may be indicative of future developments in the field of color reproduction.

**Trapping.** In color printing, trapping refers to the ability of a surface to accept ink, after an ink layer has already been deposited. In normal procedure, one plate printing follows another with 6-12 hours between printings. If too much time elapses from first plate printing to last, the ink from the first printings will become so dry and glazed that other colors will not stick or trap. In some cases, plates are prepared to eliminate under colors where possible in order to permit printing close to the paper instead of on top of other inks. This is especially true when plates are planned for high-speed wet printing; for example, in shadow areas some of the color is removed, letting the black plate carry the bulk of the ink.

**Two-plate halftones.** Where less costly means are wanted to introduce color into halftones, duotones, duographs, or duotypes may be employed. Duotypes are the simplest form, consisting of two halftone plates for letterpress produced from a black-and-white original, both of the plates being made from the same negative but etched differently. One plate is etched for detail and printed in a dark color, whereas the other is etched for a flat effect and printed in a light color. In the printing operation, the two plates are printed slightly out of register; otherwise the darker color would tend to obliterate the lighter one, because the screen angles are the same.

Duotones and duographs are similar because both are made from a black-and-white original. Two negatives are shot at different screen angles; the negative for the darker color is shot high, and that for the lighter color low or flat. Duotones are printed in complementary colors (such as red and green); or, in black and a color such as red, blue, or green. Duographs are printed in a dark and light tone of the same color. Duotones and duographs are not limited to the letterpress process, as are duotypes. There is some confusion in the use of these terms; workers in the trade may use them without regard to the specific meanings.

**Fake color work.** The manipulation of black-and-white reproduction to imitate the effect of process color work is called fake color work. Four separate negatives (all alike) may be worked on by hand to get approximate colors in the finished print. A widely used material in art and copy preparation for fake color work is the Bourges process, developed by A. R. Bourges, in which transparent films in a variety of colors and densities are used as overlays for the black-and-white art work. Even relatively unskilled workers can make effective use of the Bourges techniques; in the hands of an experienced artist, the quality of the results is good.

**Blow-ups.** Just as enlarged images are used in black-and-white, so blow-ups may be used in color

work, particularly by the lithographer. After a set of corrected negatives and screened positives has been completed for an advertisement or small folder, these positives may be enlarged for a brochure, window card, or poster, without repeating the steps of photographing through the process filters, masking, and screening. Of course, by enlarging the screened positives, the dot size (screen ruling) will become larger, but the viewing distance for the larger image will be greater, so the dot pattern will not be objectionable.

**Conversions.** Color plates prepared for letterpress may be converted to offset by pulling proofs of the plates on cellophane or thin acetate film. These positive proofs may be used directly in making deep-etch offset plates for same-size reproduction, or they may be enlarged to obtain negatives for albumen-type offset plates. Other systems of conversion from letterpress to offset which can be used for same-size reproduction are the Brightype and the d-i-Offset (direct image) method. With these systems, the entire color form would be converted, type as well as halftones. The Brightype method treats the form with a thin, black lacquer, after which the printing surface is polished with an eraser. The form is put on the bed of the special vertical copying camera and photographed to obtain a film for offset platemaking. The d-i-Offset system pulls a proof of the letterpress form directly onto a special paper-backed aluminum foil plate, which after a slight treatment is ready for the offset press.

**Flat color work.** The most obvious method of putting color into printing is that of printing one or two headlines, or display lines, in another color. To be more effective, art work in the form of sketches, borders, or backgrounds, may be printed in one or more colors, in addition to the black usually used for the main body of the text. Such colors, termed flat colors, have no variation in tone (as in process work), but appear simply as an even film of ink on the paper. A separate form, or plate, is prepared for each color, with the various impressions printed in register for the completed design. Very often the break for color is made with transparent overlays placed over the original with the required areas drawn on each overlay. The Benday method is used to break up the flat colors into patterns to give the impression of color tones. A flat tint plate is sometimes printed under a normal black-and-white halftone to add color where the expense of more elaborate handling is to be avoided.

**Posterizing.** A seldom-used but very pleasing presentation of a subject in flat colors is that known as posterizing. Here, the black-and-white photograph is copied in three (sometimes four) steps on separate pieces of high-contrast film. Exposures are regulated so the different tones of the original are interpreted as line negatives; the high-contrast film prevents individual tones from appearing, tending to blend intermediate gradations into one tone for each exposure. In using three steps, a short exposure shows up only the highlights in the

first negative; a long exposure picks up all tones except deep shadows for the second negative; and a medium exposure reproduces the middle range of tones in the third negative. When positive plates of each negative are printed in register using different tones of ink, a posterlike reproduction is the result. Dark, medium, and light tones of the same color printed on a lightly tinted paper finds general preference, but harmonizing colors may also be used.

**Split fountain.** A fairly common practice in commercial printing plants to get a variety of colors into a printed piece is the use of a split fountain. The ink fountain of the printing press is divided into several compartments, which permits two or more colors to be applied to the paper with one pass through the press. As an example, a two-color job may be run with red on one half of the paper and black on the other half. By turning the sheet around for a second impression the job can be completed with one-half the usual total number of impressions. This method cannot be used indiscriminately; the job must be planned ahead of time to take advantage of this trick of the trade.

**Fluorescence process.** Eastman Kodak introduced this method of preparing watercolor sketches and their subsequent reproduction in 1941. It is similar to the fluorographic preparation of wash drawings for highlight reproductions in black-and-white. Special pigments are furnished in kits for the artist, who proceeds to paint as he normally would. The finished work will fluoresce or glow under ultraviolet light, making possible color separations with little or no correction. Reproduction follows regular four-color process routines, but special filters are used. The copyboard is also surrounded with a special hood to allow control of lighting; either white or ultraviolet, or a mixture of both is used. Detailed instructions on the process are available from Eastman.

**Variety of modern color printing.** Color has come to be so much accepted by the American public that very little printing is produced without color in one way or another. The daily newspaper, traditionally a black-and-white medium, is beginning to use color. At the moment, only an occasional advertisement appears in color, but a few newspapers presently talk of ROP (run of the paper) color; newspapers will probably soon join magazines in offering color to their advertisers. Modern magazines have color halftones on nearly every page; color photography was first used to illustrate a story in the *American Magazine* in 1935. Textbooks in the academic range are predominantly free of color, as are novels; however, color is used for the jackets on many of them. Children's books and grade school books are filled with color, serving to make them more attractive.

The average person comes face to face with color printing in a hundred forms when he enters today's drug store, hardware store, or supermarket. The use of color in merchandising is so important that many intensive studies have been made of its ability to stimulate consumers. Besides color's effect

on visibility, it should also convey an impression of the characteristics of the contents of the package to the viewer. In this connection, color printing becomes further involved in the technicalities of getting ink on many surfaces other than paper. Cellophane, foil, waxed paper, and plastics present problems of absorption, drying, luster, and fading. Nevertheless, color printing in the field of flexible packaging and folding boxes is a major segment of the graphic arts industry.

Metal decorating, or tin printing, also has its problems with color printing. Most of this work is done by lithography, using presses which take a sheet of metal straight through between blanket and impression cylinders, without bending around the impression cylinder, as is done with paper. For process colors, a white opaque background is first printed and baked on the metal in those areas where the process colors are to appear.

There are many other specialty houses dealing in some individual way with color printing. Products, in addition to those touched upon here, include calendars, greeting cards, labels, sheet music, decalcomanias, pictorial post cards, posters, maps, and fine art reproductions. See PRINTING; PRINTING PLATE. [K.R.B.]

## Printing plate

A block or sheet of metal, wood, plastic, or other material which carries on its face a reproduction of a design, drawing, photograph, or other art work and which, when placed on a printing press and inked, transfers the reproduction to paper, cloth, metal, or any other surface capable of receiving it. In plates made for letterpress printing, the reproduction or printing image is in relief. In planographic printing such as lithography, the printing image is on the surface of the plate, on a level with the nonprinting portions. In intaglio printing (for example, rotogravure), the printing image is incised or etched below the plate surface.

Plates for planographic and intaglio printing are considered separately below under their respective headings.

**Wood engravings.** The earliest form of printing plate was the wood engraving, which is believed to have originated in China and to have been used in both China and Japan as early as the eighth century A.D. In Europe it appears first in the block books of the early fifteenth century, prior to the invention of printing from movable types and possibly earlier than 1400. The print of St. Christopher carrying the infant Christ across the swollen stream, probably the most famous of block prints, is dated 1423. Johannes Gutenberg did not make use of wood engravings, his pages being printed from type and embellished by hand, but wood-engraved initials and decorations appear in books printed by his successor, P. Schoeffer, and there are wood-engraved illustrations in books of the later fifteenth century.

In America, wood engravings, or woodcuts, were widely employed for book and magazine illustration and for commercial work as well up to the time of

the introduction of the photomechanical halftone in the early 1880s. It was from the cutting done in the making of a wood engraving that the term cut came to be applied to photoengravings generally. The industry prefers use of the terms halftone or line engraving. With the advent of the halftone, which made possible the direct reproduction of photographs and other tonal copy, the use of woodcuts rapidly declined, and the process survives today mainly as a fine arts medium. Besides wood (usually in the form of end-grain boxwood), artists and art students make use of linoleum, hard rubber, and other workable materials which, when mounted on blocks of the right height, can be printed on a press along with standard type.

In making a wood engraving, the artist first either draws his design on the face of the block or transfers it to the wood by rubbing from a drawing previously made on paper. Then, with tools known as gravers or burins that resemble small gouges or chisels (a V-pointed graver is used for incising fine lines), he cuts away those parts of the block that are not to print, leaving the lines or areas of his design in relief. Shaded effects are obtained by incising fine lines, either in parallel or crossing each other as in cross hatching. If the print is to be in more than one color, a separate block is cut for each color. The blocks are then printed successively and in register one with another.

Substantially this same process has been used for centuries in the making of Japanese prints, the effect of the colors being often enhanced by skillful manipulation of the inks and overlapping of tints.

**Steel and copperplate engraving.** Another type of engraving that is made by hand is the steel or copperplate engraving, which in its original form dates back at least to the fifteenth century. This is an intaglio plate, with the lines of the design cut into the face of the metal. In printing, the plate is inked and the surface wiped, leaving the ink only in the incised lines, from which it is transferred to paper by combined pressure and suction. The ink thus transferred, when dry, produces the raised effect seen in engraved stationery and formal announcements—an effect extensively imitated by the process known as thermography.

**Line engravings.** These, the direct descendants of wood engravings, are used in reproducing pen and ink drawings, scratchboard and pebbleboard designs, and similar material in which the lines or areas of the copy are in black on (usually) a white surface. There are no intermediate grays in line copy as there are, for example, in a photograph. The copy is placed on a copyboard in front of the engraver's camera, the distance between camera and copy is adjusted to bring the image to the specified size, and an exposure is made. On the developed negative the black lines or areas of the copy are white and transparent; the whites of the copy, both within and around the drawing, are black and opaque.

In the usual platemaking procedure, the emulsion of the negative is stripped from the backing and

squeegeed down, right side up, on a plate of glass (glass flat) along with other line negatives. The flat is then placed in contact, negative side down, with a metal plate, the surface of which has previously been coated with an emulsion sensitive to actinic (ultraviolet) light. Zinc, usually in sheets of 16-gage thickness, is the metal used for most line engravings; copper, also 16-gage, is used for engravings containing fine detail or intended for long runs. Magnesium has found increasing use in line work.

Flat and plate are next put into a photographic printing frame and exposed to a powerful light. This light, passing through the white (transparent) lines of the negative, tans the emulsion underneath, rendering it insoluble in water; where the face of the plate is shielded by the black (opaque) portions of the negative, the emulsion is not affected and remains soluble. The negative thus serves as a stencil, admitting light only to what will eventually be the printing portions of the plate.

After exposure, the face of the plate is rolled with a grease ink, or the plate is dipped in dye, and washed under running water. The ink or dye adheres to the parts that have been acted upon by the light but is washed away, along with the dissolving emulsion, from the unaffected portions, thus bringing out the printing image, which can now be inspected, and at the same time leaving exposed those parts of the metal which are later to be etched away. If necessary, the image is treated with topping powder to make it acid resistant.

In the etching process, the plate is given as many exposures to the acid (bites) as may be necessary to bring the nonprinting parts down to the proper depth (0.025–0.040 in. for newspaper plates). Nitric acid is the usual mordant if the metal is zinc, iron perchloride if it is copper. Open spaces between or around the lines in relief are deepened by routing, and any dead metal is removed in the same way.

In the final stage the plate is cut up into smaller plates corresponding to the several negatives, and the separate plates, after proofing, are delivered either in that form to be used on patent base, or mounted type-high on wood or metal bases, ready to be inserted in a type form and printed with the type.

By using a photographic positive instead of a negative in making the plate, or using a negative print as copy, the blacks and whites of the copy may be reversed, so that the lines of the copy come out white and the whites of the copy black. This type of plate is known as a positive or reverse line plate. Another type of reverse plate is obtained by turning or flopping the negative when it is placed on the glass flat, thereby reversing the copy from left to right and producing a mirror image.

**Benday plates.** The Benday (or Ben Day) process takes its name from its inventor, Benjamin Day, an American artist who brought it out about 1880. Essentially, it is a method by which shadings or flat-tone effects in the form of line, dot, grain, stipple, or other patterns can be applied mechanically to a



part or parts of a line reproduction. Although it has been largely supplanted by other methods which may be applied by artists to produce similar effects, the Benday method is still used for cartoons, in both black and white and color, and for work such as maps and diagrams in which sections or areas are to be given different shadings. The patterns, of which some 200 in all are available, are in relief on the face of Benday screens—rectangles of hardened gelatin ranging in size from 6 by 8 to 16 by 20 in.

The shadings or tints may be applied to the copy itself, to a negative or positive of the copy (the latter for reverse effects), or to the plate, the last named being the technique most commonly used. For this, the copy (usually an outline drawing) is photoprinted on the metal and the parts not to receive the shading are painted over (stopped out) with gamboge, a yellowish, viscous substance which, when dry, is soluble in water. The pattern is applied by inking the Benday screen selected with a grease ink and printing it over the face of the plate, including both gamboged and ungamboged portions. The plate is then washed, dissolving the gamboge, which carries away with it the overprinted lines or dots and leaves the pattern only on those areas where it is wanted—areas that have been indicated on the copy or on an overlay by the artist. The dots or lines of the pattern are then treated with topping powder (acid resistant) and the plate is etched in the same way as a line plate.

Of the methods employed by artists to produce Benday effects, the most widely used are the boards such as Craftint (impregnated with one or two invisible patterns) and shading sheets (a technique conceived by B. F. Hutchison and developed commercially by A. R. Bourges). In the first case, the artist makes his drawing in line on the face of the board, then paints the areas to be shaded with a liquid developer which brings out the pattern in black. With doubletone boards, which carry the equivalent of light and dark Benday patterns, a different developer is used for each pattern. Patterns must be blotted after they are brought out. The result is a line drawing complete with Benday tints or shadings. Fashion illustrations and drawings showing details of machinery are among the subjects produced by this method.

The shading sheets are transparent sheets on which Benday patterns have been printed. When an area is to be shaded, the artist cuts a piece to fit from the desired sheet and applies it over the area so that it becomes, in effect, an integral part of the copy (Fig. 1). Highlight effects may be produced by scraping off the dots, lines, or grain comprising the pattern with a knife. Shading sheets are also made with the pattern printed in opaque white for lightening (graying down) black areas or heavy lettering or type.

An advantage of both the board and the shading-sheet techniques is that the shading is applied by the artist and is entirely under his control. Both are also faster than the regular Benday process. With

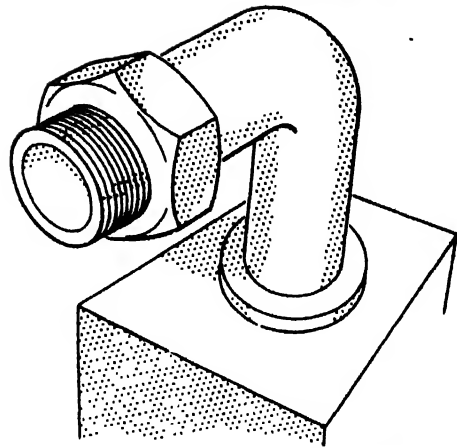


Fig. 1. Line drawing with shading sheet overlaid to produce halftone effect. Screen pattern removed from areas where solid white shows. Solid black was inked on original drawing. (American Museum of Photography)

both, the final result is line copy which may be photographed directly without use of a halftone screen and processed in the same way as an ordinary line plate.

Although it is not a shading process, mention should be made here of the Tone-Line process, introduced by Eastman Kodak, by which line reproductions may be made from photographs. A continuous-tone negative held in contact with a transparent print, much in the manner of bas-relief photography, is the basis of this process.

**Halftone plates.** These are plates made from photographs, wash drawings, and other art work done in continuous tones that range from white through intermediate grays to black—as distinguished from the all-black lines and areas of line copy. To reproduce (or simulate) these tones, use is made of the halftone screen, which in its standard form consists of two plates of glass, each ruled (or etched) with fine parallel lines, and cemented together face to face so that the lines on one cross the lines on the other at right angles. The lines are opaque, the spaces between them transparent; and lines and spaces are the same width. In the screen generally used for black and white work, the lines are at angles of 45 and 135° with the sides.

The size of a screen is known by the number of these lines to the linear inch. For the general run of letterpress printing, this ranges from 50 to 150, selection of the size to be used depending primarily on the smoothness of surface of the paper on which the halftone is to be printed. Thus, for newspaper work, screens of 50–65 and sometimes 85 lines are normal, with 65 the size most commonly used; for English or machine finish papers, 85–100 lines; for coated papers, 110, 120, and 133 lines; and for superfine papers and special deluxe work, 150 (Fig. 2). In general, within this range, the finer the screen, the more faithful the reproduction will be to the original copy. Screens of up to 400 lines are made, mostly for specialized purposes.



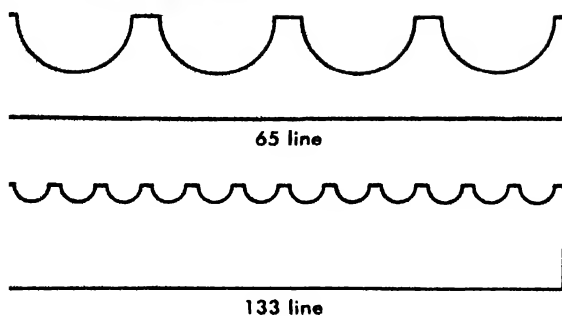


Fig. 2. Cross section of photoengraved halftones showing the relief character of dots and the variations in spacing when different screen sizes are used. (American Photoengravers Association)

The first mechanically ruled cross-line screen was produced by F. E. Ives in 1885. Prior to that, however, in 1880, the famous Shantytown halftone, the first halftone printed in an American newspaper, had appeared in the *New York Daily Graphic*. This was made by S. H. Horgan, using a one-way screen. Ruled and etched screens were introduced commercially in the middle 1880s by M. and L. Levy.

In the making of a halftone, the screen is placed in the camera directly in front of the film or photographic plate but not in contact with it. The space between, known as the screen distance, varies with the nature of the copy but ranges generally from  $\frac{1}{8}$  to  $\frac{1}{2}$  in. Copy is placed on a copyboard in front of the camera and an exposure is made.

Of the various explanations that have been made of the effect of the screen, the most acceptable would seem to be a combination of the pinhole lens and light diffraction theories originally advanced by Ives and M. Levy, respectively. Essentially, and nontechnically, the network of lines in the screen breaks up the image into dots of different sizes (Fig. 3). White areas in the copy reflect the most light through lens and screen, black areas the least, gray areas in proportion to their lightness or darkness of tone. When light from the white areas passes through the screen, each of the tiny openings is believed to serve as a pinhole lens, causing the rays to diffuse and to register on the negative as relatively large black dots. In white areas such as highlights, these dots will overlap, producing on the negative the effect of very fine white dots on a black background. Similarly, areas that are black on the copy register on the negative as pinpoint black dots, and the grays register as dots of varying intermediate sizes.

In the platemaking stage, these dots serve the same purpose as the opaque black portions of the negative in line engraving, shielding the emulsion and keeping it soluble so that it dissolves when the plate is washed, exposing the metal for subsequent etching. The surface of a halftone plate thus consists of thousands of tiny dots of different sizes which the eye, not being sufficiently microscopic to see them as dots, blends into the darks, mediums,

and lights of the original copy. The halftone effect, in other words, is actually an optical illusion.

Besides the regular network-type screens, numerous special screens are available. These include the Kodak contact screen, on film, which is placed in contact with the negative, and which has found more use in offset than in letterpress printing; grain and wavy line screens; and the double-textured Grafatone screen which combines two textures, such as 65 with 85 line and 100 with 120.

Halftone plates may be finished in several ways. Most widely used is the square (rectangular) halftone, common to all forms of printing. Others are the silhouette or outline, much used in magazine advertising, in which the dots are removed from the background, leaving the figure or product silhouetted against the white of the paper; the highlight or dropout, a favorite for fashion advertising in newspapers, in which the dots are removed from the highlights by either manual or photographic methods; and the vignette, whole or partial, in which one or more edges are feathered out or shaped into irregular backgrounds. Where type and halftone or line and halftone are combined on the same plate, the result is known as a combination plate.

**Color plates.** Relief plates for color printing are more fully considered elsewhere (see PRINTING IN COLOR). For full-color reproduction of paintings, color photographs, and colored products such as fruits or fancy merchandise when shot direct, the four-color process is generally used. In this, by means of photographic filters, color separation negatives are made for each of the three primary colors and for black. From these in turn are made halftone plates which carry the yellow, red, and blue values, with the black plate supplying definition and shading.

In making these plates, a rotatable circular screen is used by means of which the screen angle (angle of the lines of the screen with the horizon-



Fig. 3. Greatly enlarged detail of photoengraved halftone plate showing dot formation. (American Museum of Photography)

tal) is changed for each color. This is done so that the dots for the four colors, when printed, will tend to lie alongside one another, not on top of one another, thereby avoiding moiré or patterning and enabling the eye to blend them into secondary and tertiary colors. Where blue and yellow dots, for instance, fall side by side, the eye sees them as a green, and this may be a blue-green or a yellow-green according to predominance of blue or yellow dots. Similarly, reds and yellows blend to produce oranges, and blues and reds to produce violets or purples, while tertiary colors such as certain browns are produced by the blending of three and sometimes four colors. Here again the effect depends on optical illusion.

Photoengravers differ as to which screen angles give the most effective results but agree on a separation of  $30^\circ$  between colors. In what is probably the most widely used combination, the screen is set at  $45^\circ$  for the black (the angle at which the screen lines are least noticeable), at  $15^\circ$  for the blue, at  $75^\circ$  for the red, and at  $90^\circ$  for the yellow, which, being the least conspicuous color, is not noticeably affected by the half separation.

Extensive hand finishing, done mostly by reetching and (within limits) by burnishing, is generally needed on sets of color plates to lighten or darken parts of a plate or plates and to bring the colors into proper balance.

Several methods are available by which color plates may be made without the use of color separation negatives, or colors may be faked from black and white copy, but results are seldom as good as when the negatives are used.

Full-color reproduction may also be accomplished by the three-color process, in which the colors used are yellow, red (magenta), and a blue darker than the cyan of four-color process. It is claimed by some photoengravers that truer color values can be obtained by this process, but four-color is generally favored, partly for the reason that the black is available for printing type, the legibility of which is weakened by printing it in color.

Two-color plates are of many kinds and combinations. They may, for example, be made from copy prepared in the two colors that are to be reproduced, such as a painting done in tones of gray and red. Where two complementary colors can be used, such as a yellow-orange and a dark blue, a surprisingly wide range of color effects may be obtained by skillful blending of the colors. Monochromatic copy such as a photograph of a seascape may be reproduced in two colors by making two negatives, one with contrast and detail to print in black, the other a flat or gray negative to supply color values and print in blue. The latter will be made with a  $30^\circ$  difference in screen angle. Such plates are known as duotones or duographs. Tint blocks or tint plates of metal or hard rubber, with either solid or screened surface but with no design, and reverse plates of type or of line drawings are used to underprint type or illustrations, as often seen in magazine advertising.

**Electronic engravers.** Plates for letterpress printing can be produced by electronic engravers, generally from tonal copy and without the use of cameras and platemaking equipment. Use of this device has developed rapidly since the 1930s. The machines offered commercially produce halftones, varying from 50 to 200 lines/in., on plastic, zinc, magnesium, aluminum, copper, or brass according to the manufacturer. Machines of this type include the Fairchild Scanagraver, an American product; the Klischograph, developed by R. Hell of Kiel, Germany; the Elgrama, produced in Switzerland; and the Hassing engraver, invented by O. Hassing. These machines produce halftone dots similar to the traditional engraver's dots. The Hassing machine can be used to produce an unusual triangular dot pattern in addition to the usual dot structure (Fig. 4).

On the Fairchild Scanagraver, which was the pioneer American machine, the copy, usually a glossy photographic print, is bound around a small horizontal cylinder at the right of the machine. Mounted above this is an electric-eye scanner. On the left of the machine, on the same shaft, is a similar cylinder, around which is bound a thin sheet of plastic. Above this is a stylus, the point of which is electrically heated. In a separate housing between the cylinders are the controls of the machine, which govern the depth of the bite made by the stylus.

In operation, as the cylinders revolve, the scanner picks up the blacks, grays, and whites of the copy and converts them into electrical impulses of varying intensity which are transmitted to the stylus. This in turn bites into the face of the plastic, much as the acid bites into the spaces between the dots in a metal halftone. White areas receive a deep bite which leaves a small dot, black areas a shallow bite which leaves large dots, with the grays falling in between. On the 65-line machine, which

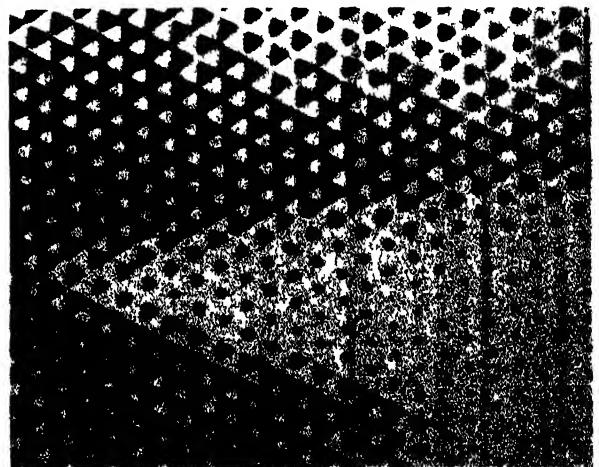


Fig. 4. Electronically engraved halftone dots. This highly enlarged detail of a halftone plate shows both dot structure in shadow and highlight areas and the special Hassing triangular dot. (American Museum of Photography)

is made for newspaper use, there are 65 such dots to each peripheral inch. At the same time the scanner and stylus, which are attached to grooved rollers and begin their action over the inner ends of their respective cylinders, move gradually outward,  $\frac{1}{16}$  in. at each revolution of the cylinders, producing the equivalent of a plate made with a 65-line screen. Plates up to 8 by 10 in. in size can be turned out in 30 min or less. Trimmed to the desired size with shears or a cutter, these may be mounted type high on blocks, to print with type, or may be attached with two-way adhesive to blank spaces left for them on the face of a stereotype plate.

Line copy such as cartoons may also be run on the Scanagraver but will come off the machine with a background of fine black dots over the face of the plate, similar to the dots of a Benday pattern.

Closely related to the electronic engraving machines are the electronic scanners of Printing Developments, Inc. (American), and Hunter-Penrose (British). In the Printing Developments scanner, four-color separation negatives, without halftone screen, are made simultaneously from flexible transparent color copy such as Anscochrome and Ektachrome films. The Hunter-Penrose equipment makes separation negatives one at a time and scans only flat or reflective copy.

**Duplicate plates.** These, as the name implies, are duplicates of original plates or type pages. They are used for long runs, as in the case of metropolitan newspapers and nationally circulated magazines; for jobs where the originals must be preserved for future use; for jobs to be run in multiple (printed from two or more identical plates or sets of plates); and when identical plates must be sent to several printers or publishers for use at the same time.

The oldest form of duplicate plate is the stereotype, used mostly by newspapers. In this process the type page is locked firmly in a heavy metal frame called a chase and a thick sheet of papier mâché forced down upon it by mechanical or hydraulic pressure so as to form a mold or mat (short for matrix) of the page. From this a plate is cast in type metal which duplicates the printing surface of the original page. If the plate is intended for use on a large rotary press, it will be semicylindrical in shape; if for use on a smaller tubular rotary press, cylindrical; and if intended to print with type on a flat-bed press, flat. The stereotyping process is quick, relatively inexpensive, and specially suited for use by newspapers publishing several editions in which the front page and certain inside pages must be remade from issue to issue.

When identical advertisements are to be run on the same date in a number of newspapers (as many as 400 sometimes in the case of national advertisers), they are usually sent to the papers in the form of stereotype mats. From these, flat-casts are made which are inserted in the type pages.

A more durable plate, used for large-circulation magazines, for much book work, and for large-edition advertising pieces, is the electrotype. For this

the plate or page to be duplicated is molded under heavy pressure in wax, tenaplate, lead, or vinylite, and the mold suspended in a bath of copper sulfate, where it is connected to the negative pole of an electric generator. A bar, usually of copper, is also suspended in the solution and is connected with the positive pole of the generator. When the current is turned on, a thin coating of copper is deposited electrolytically on the face of the mold, forming a shell which later is backed up with type metal to give the plate the strength and rigidity needed on the press. Electrotypes are usually made flat and, if intended for use on rotary presses, are curved by mechanical pressure. Flat electros, to be mounted on patent base, are normally made 11 points (0.152 in.) thick (Fig. 5).

For extremely long runs requiring maximum plate durability, electrotypes are faced with nickel or chromium. Where the former is used, the plates are sometimes mistakenly referred to as steel-faced electrotypes. Nickel, besides being more durable, is more resistant than copper to the chemicals found in certain printing inks.

Electrotypes cost more and take longer to make than stereotypes, but the process gives a more durable, better-quality plate, suitable for use on smooth-finish papers and capable of exact duplication of fine-screen halftones and process color plates.

Duplicate plates of vulcanized rubber and of plastic have found increasing use in recent years in book work and many forms of advertising and commercial printing. Advantages include light weight, ease of handling on the press, and lower shipping cost. The Dupont flexible photopolymer plastic plate, "Dycril," introduced in 1959, has many features to recommend its use, such as lightness of weight, durability, and speed of production. This plate, made of a thin layer of photosensitive plastic bonded to a metal support, can be used in relief printing on any type of letterpress equipment and can also be used for dry offset on presses

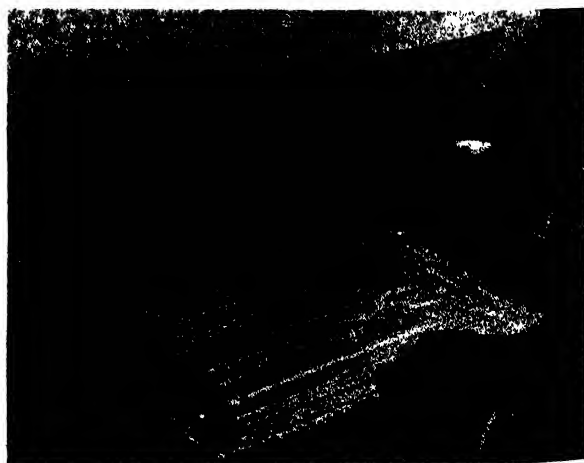


Fig. 5. Removing electrotype shell from vinylite mold. Shell will be backed up with metal to give it rigidity for printing. (Rand-McNally)

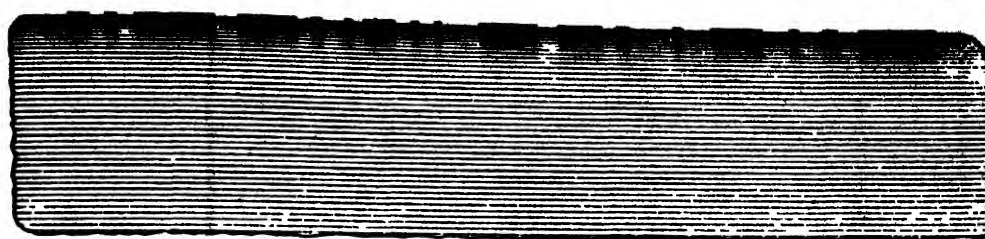


Fig. 6. Litho stone absorption.

which will accommodate a plate 0.030 in. thick. Processing time for the Dycril plate is 15–30 minutes. After contact-exposure of plate and negative to an ultraviolet light source, the plate is subjected to a pressurized water spray containing a minute amount of sodium hydroxide. This operation washes away the unexposed portion of the plate and leaves the printing areas in relief. Although not primarily intended for that purpose, photopolymer plates can also be used for making mats or molds for stereotypes and electrotypes either by cold pressing or rolling. [L.W.S.]

**Planographic plates.** Printing plates of the planographic type are intended for lithography, a method invented in 1798 by A. Senefelder and based on the mutual antipathy (repulsion) of grease and water. It was originally carried out on litho stone, a stratified secondary rock found at Solenhofen, Bavaria, in the form of compact homogeneous slabs consisting chiefly of calcium carbonate.

**Lithography.** Litho stone is almost perfectly adapted for lithography because of its porosity and natural tendency to absorb both grease and water. An illustration drawn with a greasy ink or crayon penetrates into the surface of the stone, as does a dilute solution of nitric acid and gum arabic (Fig. 6).

The drawing or illustration forms the actual printing image on the stone; the acid-gum solution acts as an etch and converts the bare surfaces of the stone into calcium nitrate, a chemical surfacing that remains moderately damp when moistened with water.

When the stone bearing the drawing is moistened with a thin aqueous solution of gum arabic and the wetted surface immediately rolled up with greasy litho printing ink, the ink adheres to the greasy drawing and is repelled by the moist (nonprinting) areas on the stone. This principle, established by Senefelder, is the basic application of all lithographic printing, including offset.

Lithography from stone is a direct procedure because impressions on paper are taken from the inked images on the stone. The method really is a fine-arts process because the quality of the reproduction depends upon the skill of the artist. The results possible with the method were shown in the lithographs depicting scenes of American life introduced in 1857 by N. Currier and J. M. Ives.

Despite its utility as an artist's medium, litho stone has several disadvantages. It is expensive,

heavy, and cumbersome, and requires that printing be done on slow flat-bed presses. Senefelder realized this and attempted in 1800 to use zinc sheets as litho printing surfaces. Aluminum plates were suggested in 1891 by J. Mullaly and L. L. Bullock, but utilization of metal plates for litho printing did not see wide practical application until introduction of the offset press.

**Offset printing.** The press designed in 1881 by F. Champenois and E. Missier for what is today known as dry offset or letterpress transfer printing was the real beginning of offset printing. Lithographic application of the offset principle can be credited to I. W. Rubel, who constructed in 1904 a press incorporating three synchronously rotating cylinders, together with two reservoirs for automatically feeding fountain solution and printing ink to the litho pressplate attached to the first cylinder. The second cylinder bore a flexible rubber blanket to which the ink impression was transferred for retransfer to a sheet of paper attached to the third cylinder.

Litho metals have little porosity, and some means must be provided for retaining a film of moisture on their surfaces. This is done by mechanically graining one side of the plate with abrasives, the grain having the dual function of acting as a moisture reservoir and providing a tooth or anchorage for the litho image on the grained surface (Fig. 7).

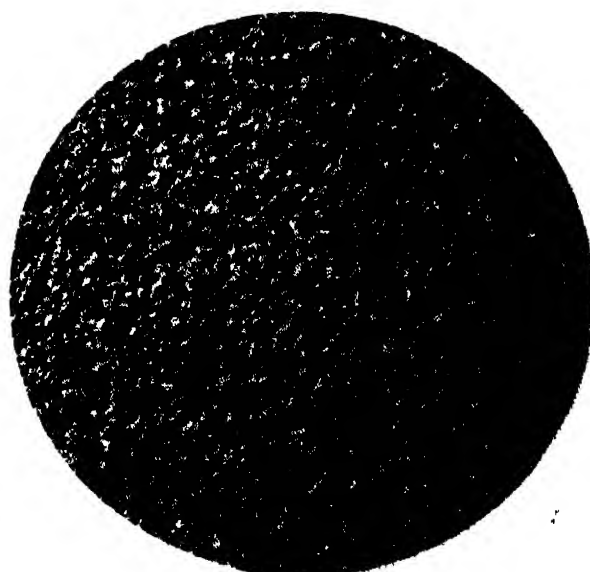
Fig. 7. Grained litho plate,  $\times 50$ .



Fig. 8. Preparation of negatives for the making of an offset plate. Here film of type is being stripped into position with film of pictures. (Rand-McNally)

Most litho plates today are produced by photolithography; an early method was the albumen process, invented in 1855 by A. L. Poitevin. It requires the grained surface to be sensitized with an aqueous solution of bichromated egg albumen. This coating becomes insoluble (tanned) through the action of actinic light.

After exposure under a line or halftone negative, the plate is covered with a thin film of greasy developing ink to promote visibility of the exposed image and impart an ink-attracting surface. The image on the inked plate is developed with tap water, which washes away the soluble (unexposed) portions of the albumen coating and leaves the inked image firmly attached to the grained surface of the metal.

Of great importance to lithographers are photo-composing (step-and-repeat) machines, originated in 1906 by W. C. Huebner. They facilitate accurate placement of images in any predetermined position on the surface of sensitized litho plates, thus rendering possible duplicate (repeat) prints which are the practical equivalent of electrotypes in a type form for letterpress printing.

When many thousands of impressions are required, deep-etch plates are likely to be used. Produced in a number of ways, such plates entail use of line and halftone positives and differ from the albumen or surface variety in that the printing image is brought into direct contact with the metal by a method of image reversal carried out chemically during the operation of platemaking.

Still greater durability is achieved with bimetallic plates, dating from experiments conducted in 1853 in France by H. Garnier and A. Salmon. Bimetallic plates consist of two different metals laminated together as an integral sheet and based on the supposition that some metals have greater affinity for litho ink.

Certain plates of this category are trimetallic—made up of three different metals or alloys, such as

zinc (or stainless steel), copper, and chromium. They are more expensive than the bimetallic variety. A popular combination for the latter type is copper and chromium. The image areas are of copper because this metal is readily rendered ink receptive, whereas chromium is considered to be more water receptive and therefore forms the non-printing areas.

More economical (although less durable) litho surfaces are plastic and presensitized plates. They are simpler to make and were originally intended for small offset presses (office duplicators) typified by Multilith and Rotaprint machines.

Plastic plates have smooth surfaces and are properly represented by articles having supports of resin-impregnated papers or saponified sheets of cellulose esters. Some of the surfaces are adapted for direct images while others are of the photolitho type and sensitized either with bichromates, silver or ferric salts, or diazo compounds.

Presensitized plates have a sensitizer incorporated in the surface coating by the manufacturer and are ready for exposure by the user. A common form of such plate is a thin sheet of aluminum coated with a diazo compound contained in a solution of polyvinyl alcohol. Diazo compounds are the preferred sensitizer because the plates have longer keeping quality (shelf life) than do surfaces sensitized with chromates.

Collotype plates are a somewhat different litho surface, used in the collotype or photogelatin process. Invented in 1855 by Poitevin, a collotype plate consists of a thin aluminum sheet having a delicate surface grain and sensitized with a coating of gelatin and potassium bichromate.

The plates are used with line and continuous-tone negatives, the exposed surfaces washed in cool running water to remove all visible traces of bichromate from the tanned gelatin images. During washing a peculiar reticulation takes place in the exposed gelatin coating and serves as a grain or tone-translation medium for reproducing the detail and gradation of the original subject (Fig. 9).

During printing, the collotype plate is kept moist with a glycerine-water mixture, and the final result is an ink impression which has the appear-



Fig. 9. Collotype grain formation.



ance of a photograph. Collotype is sometimes preferred because of beauty and fidelity of reproduction, but it is not distinguished by durability. Printing from a collotype plate is much slower than offset lithography, and the maximum number of good impressions that can normally be expected from a collotype plate is about 5000.

**Intaglio plates.** These plates differ from the planographic variety in that the printing image is incised below the surface of a metal plate. They are produced by engraving and etching.

The two terms are often confused and used synonymously, but engraving properly refers to manual inscription of designs on surfaces with the aid of gravers or special cutting implements. The term etching denotes incision of relief or intaglio designs into metal surfaces either by chemical or electrolytic action, and without using engraving tools of any kind. Etching and engraving sometimes are combined into a single procedure, such as the execution of so-called drypoint etchings.

Intaglio engraving is the oldest of all platemaking methods. Copper was used for the purpose in 1495 by the German artist Albrecht Dürer. Copper-plate engraving is a favored medium for social requisites, including personal stationery, wedding invitations, and fine work requiring only a limited number of copies.

Another form of intaglio engraving is that performed on steel plates, a method originated in the early nineteenth century. Steel engravings are more durable than those executed on copper, but the results are not as delicate because of the harder nature of steel. The most important application of intaglio engraving is for banknotes and securities, and for the production of commercial stationery in the form of business letterheads.

Copper is now a popular metal for etching and was first used for that purpose in 1520 by the Dutch artist Lucas van Leyden. It is preferred for etching procedures such as drypoint and aquatint, as well as artistic engravings such as mezzotint.

Photointaglio etching doomed manual methods of engraving and etching. It was introduced in 1826 by J. N. Niepce, who thereby invented photoengraving and pioneered in the basic principle of photomechanics.

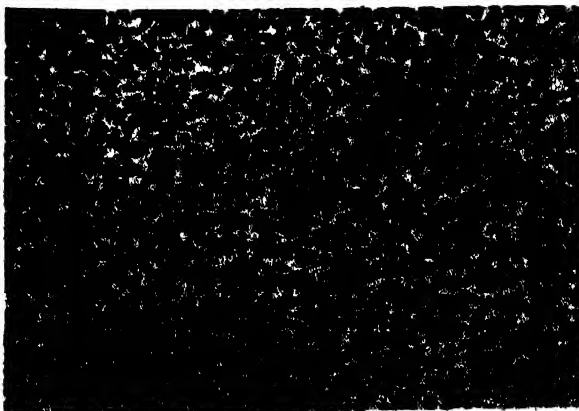


Fig. 10. Photogravure grain formation.

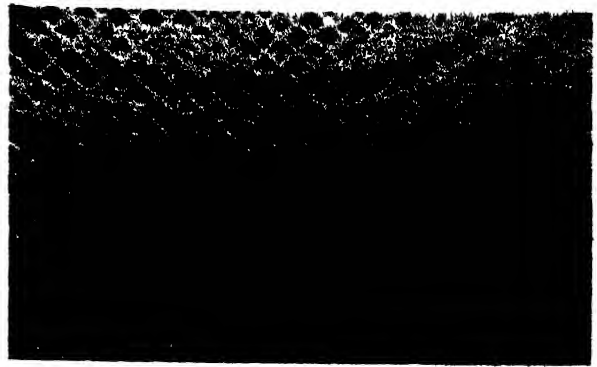


Fig. 11. Rotogravure printing surface.

There are two main methods of photointaglio etching, photogravure and rotogravure. With the first, detail and tone values are portrayed by a resinous dust grain on the surface of copper plates (Fig. 10). Rotogravure is a mechanized version of photogravure which it has displaced except for work of the highest quality. It utilizes a ruled screen for translation of tone values. The screen divides the intaglio etching into cells (depressions) of uniform size but of varying depth according to the tone values of the original (Fig. 11).

Both photogravure (1877) and rotogravure (1890) were invented by one man, the Bohemian artist and photomechanical researcher K. Klic. Rotogravure can be done either on flat copper plates or copper-surfaced cylinders, and impressions can be taken from the etched surfaces on either sheet-fed or rotary (web-fed) presses. The greatest application of the process is in publication printing and assignments involving very long press runs of millions of impressions.

The only common denominators between photogravure and rotogravure are that both methods use line and continuous-tone positives and entail photoresists in the form of variably insolubilized negative images produced by exposure of the positives on carbon tissue (pigment paper) sensitized with a coating of bichromated gelatin impregnated with a red-colored pigment. Etching of the images on the copper surfaces is performed with a series of ferric chloride solutions of progressively weaker strength (43–35° Baumé).

Inking of photogravure plates is frequently performed by hand. In rotogravure, the surface of the etched cylinder is flooded with a fluid ink of high volatility (drying power) and the excess ink is wiped from the surface of the cylinder with a flexible steel doctor blade before the inked cylinder is brought into contact with the web of paper traveling at high speed through a rotary gravure press. See PRINTING; PRINTING PRESS.

[J.S.M.E.]

**Bibliography:** H. M. Cartwright and R. MacKay, *Rotogravure*, 1956; L. Flader and J. S. Mertle, *Modern Photoengraving*, 1948; J. S. Mertle, *Evolution of Rotogravure*, 1957; J. S. Mertle and G. L. Monsen, *Photomechanics and Printing*, 1957; L. W. Siple, *A Half Century of Color*, 1951; L. W. Siple,



*The Photomechanical Halftone*, 1958; V. Strauss, *Lithographers Manual*, 1958; B. E. Tory, *Photolithography*, 1953.

## Printing press

Several kinds of printing presses are used in the three methods of reproduction most widely employed in the graphic arts, namely, relief (letterpress) printing, planographic printing (as in lithography), and intaglio printing (steel and copperplate engraving, and gravure).

Letterpress presses print from a relief surface, such as printers' cast type (Fig. 1a).

In lithography (both stone lithography, the original hand method, and offset lithography, the modern photographic method) the printing is from a planographic (even) surface, using the principle of affinity of ink and grease for printing areas, and the repellent qualities of water and grease for the nonprinting areas of the lithographic printing plate. Stone lithography prints from the stone directly on the paper sheet. On offset lithographic presses the printing is from a metal plate to a cylinder surfaced with a rubber blanket, which offsets the impression to the sheet of paper, tin, or other substance to be printed. In dry offset, a relief-etched printing plate 0.025 in. deep is used to print the impression on the rubber blanket, which offsets the impression to the sheet as in regular offset. The relief-etched plate eliminates the need for water-repellent facilities (hence, the name dry offset). In direct lithography, a thin metal plate is wrapped around a cylinder (the same as in offset and dry offset) and the impression is printed directly to the sheet, instead of being offset to a rubber blanket and then to the paper. Coarse grade work, such as posters, is generally produced by this process (Fig. 1b).

In intaglio, the printing is from ink deposited below the plate surface (Fig. 1c), as in hand-engraved or etched images in copperplate printing, steel-die stamping, or photogravure (sheet-printing production) and rotogravure (web-printing production from rolls of paper stock).

**Printing plates.** Printing plate requirements vary for printing presses in each of the three basic methods of printing. Letterpresses print from original relief plates, or forms consisting of type, photoengravings, or both, in either line plates or halftones, locked in a chase; or from reproductive plates, such as stereotypes, electrotypes, rubber, or plastic

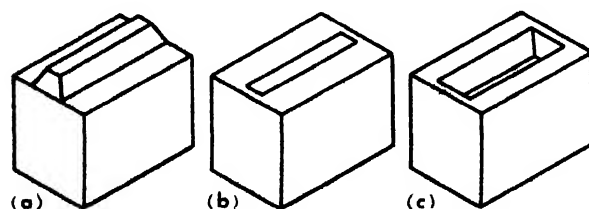


Fig. 1. Surfaces used in printing. (a) Relief, above the surface. (b) Planographic, on the surface. (c) Intaglio, below the surface.

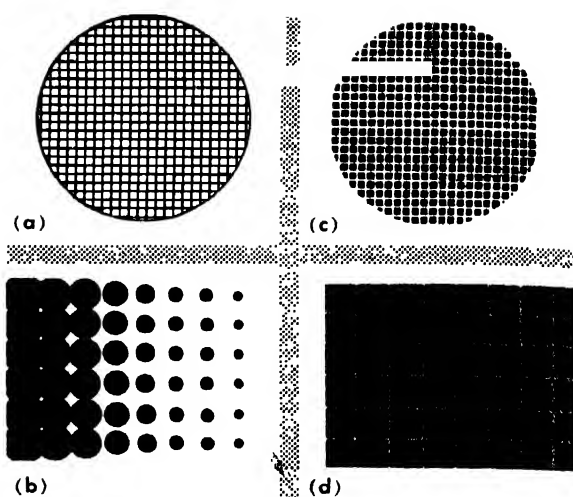


Fig. 2. Printing plate screens. (a) Halftone screen with black lines and white dots. (b) Halftone screen with variable dot areas and equal ink-film thickness. (c) Rotogravure screen with white lines and black dots. (d) Rotogravure screen with equal dot areas and variable ink-film thickness.

printing plates made from the original type, and photoengraved plates. The line-plate photoengraving consists of variable line thicknesses and shapes for tonal valuations in an illustration. The halftone photoengraving reproduces tonal pictures or continuous tone photographs through a system of graduated dots in relief (Fig. 2a and b).

The halftone process produces halftone plates in which the gradation of tone in the photograph is reproduced by a system of graduated dots produced by a screen in which a network of fine lines that cross each other at right angles is placed between the camera lens and the negative. The halftone printing plate surface consists of dots of various sizes uniformly placed, and is capable of reproducing the highlights and shadows and all the gradations of tone in a continuous tone photograph. The screen used to make the plate is designated by a number, which indicates the number of lines (both ways) to 1 in. Screen numbers vary, but the most frequently used are 65, 85, 100, 120, 133, and 150 lines/in. in both directions. In a 65-line screen plate, there are  $65 \times 65$ , or 4225 halftone dots/in.<sup>2</sup> In a 150-line screen, there are  $150 \times 150$ , or 22,500 dots/in.<sup>2</sup> In relief and lithographic printing, the halftone dot uses the principle of variable dot areas and equal ink film thickness to reproduce a halftone illustration. In letterpress printing, this ink film thickness is approximately 0.0002 in. In offset lithography, the ink film thickness is approximately 0.0001 in.

Photogravure and rotogravure halftones in monochrome (black, sepia, or other single color) are normally reproduced in dots of equal area and uniform placement but varying in the depth to which they are etched in the copper plate or cylinder. Gradation of tonal values is controlled by the amount of ink deposited on the sheet from each

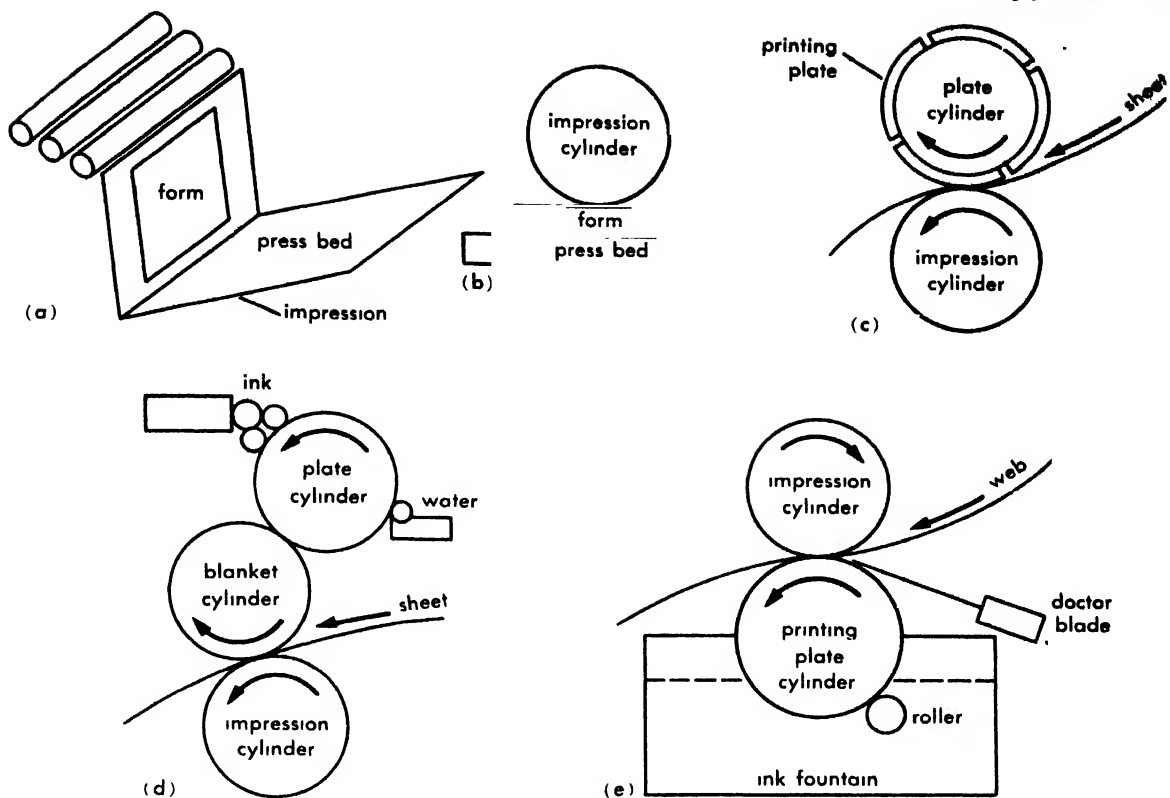


Fig. 3. Printing press principles. (a) Platen. (b) Flat-bed. (c) Rotary. (d) Offset lithographic. (e) Rotogra-

"well" or dot, this in turn being governed by the depth of the well or dot. The process thus uses the principle of equal dot areas and variable ink film thickness. The screen lines are retained in the printing plate or cylinder, the dots being etched out between them, and serve to support the doctor blade, which wipes the cylinder clean and free from superfluous ink, leaving ink only in the etched dots, before the impression is made on the paper (Fig. 2c and d). The screen generally used for monochrome work in gravure is known as the crossline screen and consists of white lines with opaque spaces between, the spaces being wider than the lines. For color work, a process of comparatively recent development is now quite widely used—the Dultgen Halftone Intaglio Process—which makes it possible to vary the size as well as the depth of the dots, thus combining the variable area and variable depth principles. See PRINTING PLATE.

The silk-screen process of reproduction differs from the relief, planographic, and intaglio methods in that the impression is pulled through a special cloth screen (or wire screen in ceramic printing), containing the image to be printed. A hand cut (with a glue base) or photographic (light-sensitized base) image is placed on a prepared screen, with the cleaned-out portion to carry the ink through the screen, and is stretched across a frame. The ink is poured into the frame at one end, and carried across the screen with a rubber-edged squeegee, which pushes the ink through the screen

and onto the sheet which is in contact with the screen during the operation. The silk-screen process (sometimes called paint-screen process) is similar to the mimeograph process used in office reproductions of letters, where the message is typed on a typewriter through a stencil, and the message is printed on a mimeograph machine built to squeegee the ink through the stencil stretched around the machine cylinder (Fig. 3a, b, and c).

**Relief (letterpress) presses.** Relief printing presses are mostly of three general types: (1) the platen press, in which both the printing form and the press bed (platen) are flat or plane surfaces; (2) the flat-bed press, in which the printing form is flat and the impression is pulled against a cylinder (one flat and one curved surface); and (3) the rotary press, in which the printing form is a curved plate form, and the impression is pulled against a cylinder (two curved surfaces).

Press make-ready is the operation of eliminating the imperfections in materials and press in preparation for running the job so that the printing plates and type will present an even surface to the paper. The order of increasing difficulty in make-ready is rotary press, flat-bed, and platen press. Platen presses are the most difficult to make-ready because the two flat surfaces of the form and the platen (or press bed) present the problem of equalizing the impression of the entire form over the entire area of the job. Flat-bed presses, with one flat and one curved surface, print a part of the form in

a continuous line, the width of which is controlled by the size of the impression cylinder. Rotary presses, with two curved surfaces of the same size, print the form in a continuous line thinner than that of the flat-bed presses. Rotary presses with large impression cylinders, printing four and five colors from smaller plate cylinders, have a wider pitch line (width of line printed at one time) than the smaller cylinders.

**Platen presses.** These presses, also known as job presses because of the ease of hand-feeding the sheets to the press, have been used for many years in educational training activities. The platen press is used for small-size jobs that can be hand fed, as well as for heavy card stock jobs that cannot be fed into a press which uses a curved cylinder to pull the impression.

High-speed platen presses have automatic mechanical feeders built into the press; such feeders have increased the production of this type of press and put it in competition with the automatic jobbers (small cylinder presses). The addition of the automatic feeder has made it possible to produce heavy card stock jobs at relatively high speeds. The clamshell action of such presses makes them especially suited to embossing and stamping operations.

**Flat-bed presses.** Large flat-bed presses are also known as cylinder presses, and the smaller sizes as automatic job cylinder presses. The automatic job cylinder presses, with high-speed operation, successfully compete with the slower-speed, large-size cylinder presses. On most job cylinder presses, the flat bed is in a horizontal position to provide for sliding the form on the press for lock-up and production. Some job cylinder presses have their beds in a vertical position in the press. These require locked-up forms to be lifted into the upright position and snapped into place by a spring lock.

Cylinder presses, both large and small, are generally of two-revolution design, one revolution to print the sheet and one revolution to deliver the sheet. These presses have small-size cylinder circumferences, as compared with one-revolution presses. The one-revolution large-circumference cylinder presses print and deliver the sheets in one revolution of the cylinder.

Automatic job cylinder presses have built-in mechanical feeders and pile deliveries. They are similar in feeding, inking mechanism, and delivery to the large cylinder presses. They run smaller-size sheets at greater press speeds. They are now being built for one- and two-color printing. Two-color and perfecting presses are similar in construction to single-color presses except that auxiliary impression units are mounted in the press frame.

Flat-beds, or large cylinder presses, are built to produce one and two colors on one side of the sheet. Automatic feeding and pile-sheet deliveries are used on these presses.

Perfecting presses are built to print on both sides of the sheet from two forms each time the sheet passes through the press. These presses are similar to the two-color flat-bed cylinder presses,

but print both sides of the sheet, instead of two colors on one side of the sheet. Their main use is in book manufacturing printing, where each sheet is completed in one pass through the press.

**Rotary presses.** Rotary presses print from curved plates against a cylinder, or curved bed.

Sheet-fed rotary presses automatically feed and print sheets, which are delivered in a pile. These presses print from one to four and five colors at one time on one side of the sheet. Some presses print all four or five plate cylinders against one large impression cylinder, whereas other four and five-color rotary sheet-fed presses print each plate cylinder against one individual-impression cylinder, thereby using one plate cylinder and one impression (packing) cylinder for each color.

In web-fed rotary presses the paper is fed to the press from a continuous paper roll or web. Operation of these presses is similar to that of the sheet-fed rotary presses in plate and packing cylinder operations. They print multiple colors on one or both sides of the paper web. When these presses are used in book work or for newspaper production, they deliver the job completely folded into signatures of 4, 8, 16, or 32 pages.

Heat-set inks are used on high-speed web-fed rotary presses which produce at speeds of up to 1000 running ft/min. At these speeds, when four colors are being printed, it is necessary for the ink to be completely dry when the web passes through the folding delivery of the press. Paper webs printed with heat-set inks are passed through heated ovens to dry the ink.

**Aniline presses.** The new name for aniline printing is flexography. This method uses relief printing, generally with rubber plates which are fastened to a cylinder and inked by a single inking roller. The inking roller is supplied with aniline ink from two rollers in the ink fountain. The amount of ink supplied to the form roller is controlled by the spacing of the roller in the ink fountain. The presses generally are of the roll-feed web-press type and are run roll-to-roll, rewinding the web. They may also be run roll-to-sheet delivery. The cut-off operation may be adjusted to variable sizes. The presses print multiple colors, and the ink's rapid drying permits rewinding if desired.

**Special-purpose presses.** Hard-packing web presses are used for commercial production. These presses are made ready in the same manner as the other presses used in the letterpress printing industry, using a hard packing on the packing cylinder. This gives a superior printed product with a clear, sharp impression.

Newspaper web presses use newsprint paper and soft blanket impression cylinders. They do not use heat-set inks or heated ovens in production. Newsprint paper, with its absorbing qualities, used with news ink, dries satisfactorily at newspaper production speeds.

Proof presses are used for pulling all kinds of proofs used in printing production, such as galley proofs of type for first readings and page proofs for final readings. Reproduction proofs, for photo-

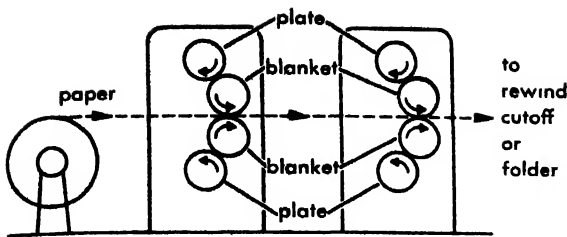


Fig. 4. Offset perfecting press. Web is threaded for running two colors on both sides of one web.

graphic reproduction of type for lithographic, roto-gravure, or other processes of printing, are also pulled on proof presses. Proof presses are likewise used to pull proofs on acetate sheets for positives in deep-etch offset and roto-gravure platemaking.

Multigraph press production is a relief (letter-press) method which uses relief-cast short type 0.095 in. high, as differentiated from type-high printers' type of 0.918 in. The short type is slid into the grooves of the Multigraph segment (a cylinder to hold the type) and the cylinder is then placed in the Multigraph machine, where the type is inked and the impression is taken on the sheet as it is fed through the press between the type and the platen. (Platen is the name given the rubber roller against which the impression is pulled.) Electrotypes and rubber plates 0.095 in. high are also used for printing on this machine.

Special-built letterpresses are designed to print specialty products such as office forms, labels, tags, and the like. Special plates of type, rubber, stereotypes, or electrotypes are used on these presses. When rubber printing plates can be used exclusively, it is possible to build a roll-fed web press with a printing cylinder of variable circumference, consisting of half shells. Two half shells make a complete plate cylinder for rubber plates; the impression is pulled against a smooth steel impression cylinder. The impression cylinder position is adjustable to the various diameter sizes of the plate cylinder used in the printing. The inking mechanism is also adjustable to the various plate cylinder sizes. The different circumference plate cylinders provide variable sheet sizes. The press cutoff adjustments deliver variable sheet sizes.

**Offset presses.** The planographic method of printing includes production on offset lithographic presses, as well as on direct lithographic presses (Fig. 3d).

Sheet-fed offset presses have a thin metal printing plate 0.025 in. thick, on which the work image is developed. These plates are made of zinc, aluminum, or copper, or a combination of two metals for the bimetal plates used for high-quality work. Plastic-coated presensitized paper plates may be used on small-size presses. The plate is clamped around the plate cylinder and dampened by passing under the water mechanism where the nonprinting parts of the plate accept moisture. It is inked by passing under the inking mechanism where the printing areas of the plate are inked. The plate prints the inked image on the surface of the rub-

ber blanket fastened to the blanket cylinder in the press. The printed image is then offset or transferred to the sheet, as it passes through the press, by pressure applied to the impression cylinder in the press. After it is printed, the sheet is passed on to the delivery of the press. Automatic feeders are used in the operation of the sheet-fed offset press.

One-color offset sheet-fed presses used in the industry vary in sheet size from 10 by 14 to 52 by 76 in. Two-, three-, and four-color presses are also available.

Offset perfecting presses print both sides of the sheet at one time, using the blanket-to-blanket method of transferring the printed impression to the sheet. This method eliminates the impression cylinder, and uses the blanket of the opposite side of the sheet as the impression cylinder to transfer the image to the sheet. Roll-fed offset web presses sometimes use this blanket-to-blanket method of printing both sides of the web (Fig. 4).

Roll-fed offset web presses are used in the same manner as letterpress web presses. Some presses print roll-to-roll (rewind), others roll-to-sheet, or roll-to-folder where complete folded signatures of a multiple number of pages are delivered.

Multiple-color offset presses are of two designs. Some transfer several colors of the printed image to the rubber blanket, and then transfer the multi-color image to the sheet in one impression. Others transfer the printed image to the sheet one color at a time, as do one-color presses (Fig. 5).

Dry offset presses operate in the same manner as wet offset presses. The plates used are etched 0.025 in. in relief, instead of being planographic (on the surface). The water mechanism is moved back from the plate contact and made inoperative during the printing operation. The inking mechanism is operated in the same manner as for wet offset, except that the rollers require a more delicate setting to prevent their inking the nonprinting parts of the relief-etched printing plate. The impression is pulled on the rubber blanket and transferred to

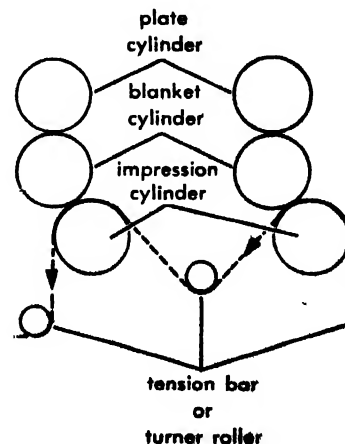


Fig. 5. Offset unit-type press for printing multiple color. Web is threaded for running two colors on one side of one web.

the sheet, in the same manner as in wet offset printing. The dry offset printing plate is really a relief printing plate, although printed on an offset press.

Direct lithographic presses operate in much the same manner as offset lithographic presses, except that the printing plate impression is printed directly on the sheet, instead of offsetting on the rubber blanket and then being transferred to the sheet; hence the name direct lithography.

In direct lithography, the printing plates read from right to left, as in letterpress printing plates and type, whereas offset lithographic printing plates read from left to right.

Direct lithographic presses can operate as sheet-fed or roll-fed web presses as desired, according to the demands of the product being printed.

**Intaglio process printing presses.** Copperplate printing presses come under two classifications, hand and power driven. The hand-power press consists of a flat bed and a curved impression surface, much like those on a flat-bed cylinder press. Great pressure is required for printing, because the sheet is forced down into the engraved plate to take out the ink. When the plate to be printed has been inked, it is wiped clean with soft cheesecloth and the printer's hand, on which whiting (a fine powder) has been applied. The plate is then placed on the bed of the press, the sheet to be printed is placed on the plate, and the impression is taken by pulling the spokes attached to the shaft of the curved impression cylinder. After the impression is pulled, the press returns to the original position automatically and is ready for the next impression.

The power-driven copperplate printing press produces by hand-feeding the sheet or card into the press for the impression. This is taken from a printing plate that has been automatically inked and wiped clean and free from superfluous ink just before the impression is taken on the sheet.

The operation of steel-plate engraving presses is similar to the operation of a power-driven copperplate press. The steel plate is used to withstand the tremendous impression pressure needed to take the ink out of the engraving on the plate during the impression cycle. Steel plates are sometimes called steel dies, the name given to the smaller size plates which are about  $\frac{1}{8}$ – $\frac{1}{4}$  in. thick. The designs are hand cut, rolled on, or etched in the steel plate for bonds, bank notes, and business stationery. The steel is given a hardening treatment to help it withstand the wear of production.

The manner of printing a steel plate is the same as that for the copper plate. The plate is automatically inked and wiped clean with a paper wipe, a roll of paper fed into the press across the steel plate and rewound during the cleaning of the plate.

The sheets or cards to be printed are fed into and removed from the press by hand. Modern steel-plate and die-stamping presses are roll-fed, with sheet delivery. Letterheads so produced are cut from the roll after being printed, and delivered in sheets. Drying of the heavy application of ink is done by infrared radiation. A specially prepared ink is used.

Steel die hand-stamping presses are small, screw-activated, swivel-operated machines for use on small dies such as monograms, crests, and trademarks.

**Rotogravure presses.** Photogravure is any of the various processes of producing prints from an intaglio plate prepared by photographic methods. Rotogravure is a process of photogravure, or intaglio printing, in which the impression is obtained from etchings made on a copper cylinder which revolves in ink. Photogravure is classified by the industry as printing from a sheet of copper, rotogravure as printing from a copper cylinder.

Photogravure presses were originally built to print from flat sheets of copper in the same general manner as a flat-bed letterpress cylinder press. The inking mechanism and wiping mechanism were cumbersome. Rotary photogravure presses are built with a cylinder on which is fastened the etched copper printing plate with a doctor blade, a thin steel blade, to wipe the superfluous ink from the surface of the plate, which revolves in a fountain of thin ink, and with a rubber impression roller of suitable durometer to pull the ink out of the etched wells in the plate and onto the sheet being printed. The feeding and delivery parts of the press operate in a fashion similar to that of rotary presses in offset and letterpress printing.

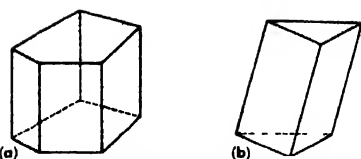
Rotogravure roll-feed web presses print from metal-coated or copper cylinders, on which the image has been etched intaglio. The cylinders revolve in ink fountains. Thin ink is rolled into the etched cylinder in the fountain. A steel doctor blade wipes off the superfluous ink, which drops back into the fountain, and the impression is taken against a rubber impression roller of proper durometer to print the image sharp and clear (Fig. 3e).

These presses run multiple colors, four on each side of a web if desired. The ink dries quickly enough to deliver the web into a folder if desired. The web may be delivered into a sheeter to be cut into sheets and delivered from the press.

**Screen process presses.** Although a considerable portion of short-run screen process (silk-screen or stencil) printing is still done with hand-operated equipment, semiautomatic platen presses with speeds up to 1200 impressions per hour and automatic cylinder presses with speeds ranging from 1200 to 3500 impressions per hour are widely used. Such presses range in size from 15 in. by 22 in. to 50 in. by 80 in. Some have been designed to print products up to 6 in. in thickness, and special presses are available which will print curved or irregularly shaped products such as bottles, cones, barrels, and cans. Because of the problem of drying, special conveyor-belt and oven arrangements are usually attached to automatic presses. *See* INK; PRINTING. [F.W.HO.]

## Prism

A polyhedron of which two faces are congruent polygons in parallel planes, and the other faces are parallelograms. The bases *B* are the congruent polygons, the lateral faces are the parallelo-



Prisms. (a) Right prism. (b) Oblique prism.

grams, the lateral edges are the edges not lying in the bases, and the perpendicular distance between the bases is the altitude  $l$ . Sections parallel to the bases are congruent to the bases. A prism is a right prism if its lateral edges are perpendicular to the bases, an oblique prism otherwise. A prism is called a triangular prism if its bases are triangles, a pentagonal prism if its bases are pentagons, and a parallelepiped if its bases are parallelograms. The volume of any prism is equal to the area of its base times its altitude ( $V = Bh$ ). See POLYHEDRON; PRISMATOID AND PRISMOID. [J.S.F.]

## Prism, optical

An optical system consisting of two or more usually plane surfaces of a transparent solid at an angle with each other. Prisms are used for deviating light. Since the amount of deviation depends on the refractive index of the prism, which varies with wavelength, prisms can also be used for dispersing light. See DISPERSION (RADIATION); REFRACTION OF WAVES.

**Reflecting prisms.** Prisms can be used in lieu of mirrors for deviating light, with the added advantage that the reflecting surfaces are protected against corrosion. In this case, there is at least one internal reflection. When the angles of incidence and emergence are zero, there is no dispersion. The over-all dispersion is also zero when the geometry of the prism is such that the dispersion at the entering surface is compensated by dispersion in the opposite sense at the emergent surface. For a detailed discussion of important types of reflecting prisms, see MIRROR OPTICS.

**Dispersing prisms.** Dispersing prisms deviate light of different wavelengths by different amounts, and they can therefore be used to separate white light into its monochromatic parts. A parallel beam of light entering the prism leaves the prism as a

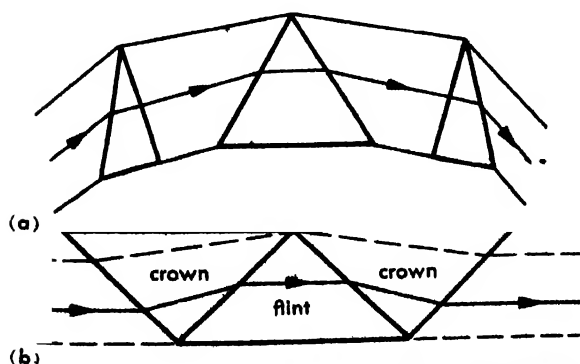


Fig. 1. Some types of dispersing prisms. (a) Rayleigh prism system. (b) Amici direct-vision system consisting of a flint-glass prism and two crown-glass prisms.

parallel beam of light but its diameter may be changed. The ratio of its diameter after refraction to its diameter before refraction can be considered as the magnification of the prism.

The prism magnification for a bundle parallel to the prism edge is always equal to unity, whereas it varies in the meridional plane (normal to the edge) as the angle of incidence is varied. It is equal to unity in this plane only if the prism is traversed at minimum deviation. If  $\alpha$  is the prism angle,  $n$  the refractive index, and  $\delta$  the deviation, it can be shown that, for minimum deviation,

$$\sin (\delta + \alpha) / 2 = n \sin \alpha / 2$$

To increase the dispersion, several prisms with their refracting edges parallel can be used. The Rayleigh prism, shown in Fig. 1a, is an example of such a system. By using a prism made of a material, such as flint glass, that has a high dispersion, and adding one or more prisms made of a material having a low dispersion, such as crown glass (see Fig. 1b), the deviation can be neutralized without neutralizing the dispersion to give a direct-vision

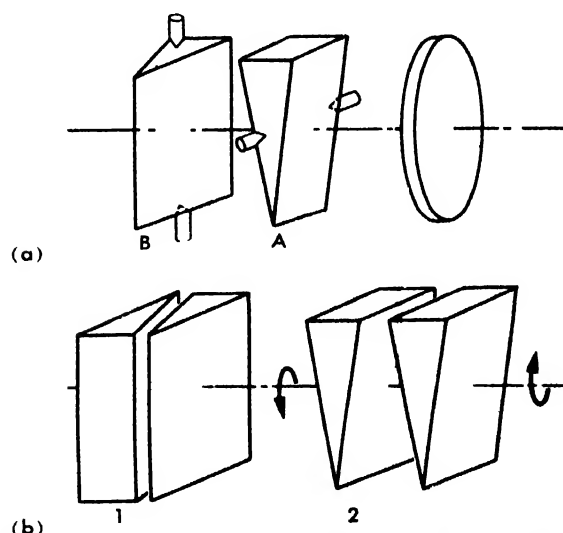


Fig. 2. (a) Pair of prisms used for varying magnification, as in zoom system. (b) Pair of prisms (Risley prism system) used for varying deviation.

prism system. The arrangement shown in Fig. 1b is known as the Amici prism system. By using a similar arrangement but adjusting the angle so that the dispersion, but not the deviation, is neutralized, it is possible to make a prism system that is achromatic over a small part of the spectrum, like an achromatic lens. See LENS, OPTICAL.

An achromatic prism in front of an optical system with its refracting edge normal to the meridional plane can be used to change the magnification of the optical system in that plane. This amount can be varied by rotating prism A in Fig. 2a about an axis normal to the meridional plane. A second achromatic prism, B, with its edge parallel to the meridional plane, can be used to adjust the sagittal magnification of the optical system. Therefore, an arrangement of two such prisms with their motions,



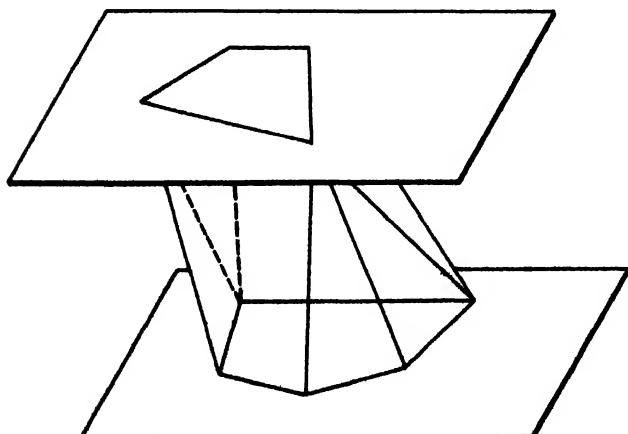
linked together can be used to form a variable-focal-length lens system called a zoom lens. See ZOOM LENS.

A thin prism is one whose angle is so small that the angle in radians is practically equal to the tangent. Such prisms are used in ophthalmology, and their powers are usually expressed in prism diopters (see DIOPTR). The Risley prism system, used for testing ocular convergence, consists of two thin prisms mounted so that they can be rotated simultaneously in opposite directions, as shown in Fig. 2b. When they are in the orientation sketched at 1, their combined deviation is zero; when both have been rotated by 90° in opposite directions, as shown at 2, their combined deviation is a maximum; at intermediate positions, their combined deviation lies between zero and the maximum, but the plane of deviation is constant. A similar pair of rotating wedges is used in certain types of rangefinders. See RANGEFINDER, OPTICAL. See also BINOCULARS; OPTICAL MATERIALS; OPTICS, GEOMETRICAL; PERISCOPE; RESOLVING POWER (OPTICS). [M.H.]

**Bibliography:** D. H. Jacobs, *Fundamentals of Optical Engineering*, 1943; H. Kayser, *Handbuch der Spectroscopie*, vol. 1, 1901.

## Prismatoid and prismoid

A prismatoid is a polyhedron all of whose vertices lie in two parallel base planes. If the two base polygons have the same number of sides, the prismatoid is called a prismoid. Among the prismatoids are included pyramids, frustums of pyramids, wedges, parallelepipeds, and other prisms. The area of a section of a prismatoid by a plane parallel to and between its bases is a quadratic function (possibly linear or constant) of the distance from the plane to one of the bases. A generalized prismoid is any solid for which the area of a section parallel to a fixed base is represented by a polynomial of degree less than or equal to 3 in terms of the distance from that base. Frustums of cones, segments of spheres, and many other solids satisfy this condition. The volume of any generalized prismoid can be expressed exactly in terms of the altitude  $h$ , and



A prismatoid.

the areas  $L$ ,  $M$ , and  $U$  of its lower base, midsection, and upper base by the important prismoidal volume formula:

$$V = \frac{1}{6} h(L + 4M + U)$$

See PARALLELEPIPED; POLYHEDRON; PRISM; PYRAMID AND FRUSTUM. [J.S.F.]

## Privacy systems (scrambling)

Devices and methods for ensuring the privacy of overseas telephone conversations handled by radio links. Such privacy can be accomplished in a variety of ways; two are particularly widely used at circuit terminals.

The first method employs equipment for inverting speech to make overseas telephone conversations unintelligible to the casual listener. At the originating end, devices change the speech-conveying signal, making high frequencies low and low frequencies high, for example. At the distant end, synchronized equipment restores the inverted speech to its original form for delivery to the listener.

A more complicated method employs filters to separate the transmitted speech signal into several narrow-frequency bands. Each of these is then treated differently in a prearranged manner, the relative positions of individual portions of the signal being interchanged in the frequency spectrum by inversion and transposition. At the distant end, associated equipment puts the full signal back together in proper shape so that the listener may receive intelligible speech. This method, particularly if such changes are automatically varied every few seconds, makes possible a radio transmission system that is very difficult to decipher. See FILTER, ELECTRIC; TELEPHONY. [C.C.DU.]

## Probability

Although probability theory derives its notion and terminology from intuition, a vague statement such as "John will probably come" is as remote from it as the statement "John is forceful and energetic" is remote from mechanics. Probability theory constructs abstract models, mostly of a qualitative nature, and only experience can show whether these reasonably describe laws of nature or life. As always in mathematics, only logical relations and implications enter the theory, and the notion of probability is just as undefinable (and as intuitive) as are the notions of point, line, or mass. An actual assignment of numerical probabilities is frequently unnecessary or impossible. For example, telephone exchanges are based on a theoretical comparison of several possible systems; only the optimal ones are built and the others discarded. Thus a huge industry depends on theoretical models of exchanges which will never exist. A simple illustration of the nature of probability models is found in Lord Rutherford's experiment.

**Example (a).** To measure radioactive intensity, Lord Rutherford proceeded as follows. Observers  $A_1$  and  $A_2$  counted scintillations on a screen and observed, respectively,  $N_1$  and  $N_2$  scintillations; of these,  $N_{12}$  were common to both observers. To estimate the unknown true number  $X$ , Rutherford assumed that each scintillation has fixed probabilities  $p_1$  and  $p_2$  to be observed by  $A_1$  and  $A_2$ , and furthermore, that the observations are independent in the sense that a scintillation observed by  $A_1$  has still probability  $p_2$  to be observed by  $A_2$ . In reality the likelihood of observing a scintillation varies with growing fatigue and the proximity of the preceding scintillation; also, the observers are affected by common causes and are therefore not independent. Equating probabilities with observed frequencies (another approximation), Rutherford sets  $N_1 = Xp_1$ ,  $N_2 = Xp_2$ ,  $N_{12} = Xp_1p_2$ , whence  $X = N_1N_2/N_{12}$ . The three equations may be solved for  $p_1$  and  $p_2$ , but these "probabilities" are purely fictitious and, as experience shows, inaccessible to experimental verification. The model is justified by plausibility and success.

**The sample space.** One speaks of probabilities only in connection with conceptual (not necessarily performable) experiments and must first define the possible outcomes. Thus, by convention, tossing coins results in heads  $H$  or tails  $T$ , regardless of experimental or philosophical difficulties the age of a person is taken as an exact number and each positive number is taken as a possible age. Throwing two dice results in one of the 36 combinations (1.1), (1.2), . . . , (6.6). An outcome such as "sum 4" is a compound event which can be further decomposed by enumeration: sum 4 occurs if the outcome is (1.3), (2.2), or (3.1). Thus it is necessary to distinguish between elementary (indivisible) and compound outcomes or events. Each elementary outcome is called sample point; their aggregate is the sample space. The conceptual experiment is defined by the sample space, and it must be introduced and established at the outset.

**Examples (b).** The experiment "distributing 3 balls in 3 cells" has 27 possible outcomes (sample points) listed in tabulation (1).

- |              |              |              |
|--------------|--------------|--------------|
| 1. {abc — —} | 10. {a bc —} | 19. {— a bc} |
| 2. {— abc —} | 11. {b ac —} | 20. {— b ac} |
| 3. {— — abc} | 12. {c ab —} | 21. {— c ab} |
| 4. {ab c —}  | 13. {a — bc} | 22. {a b c}  |
| 5. {ac b —}  | 14. {b — ac} | 23. {a c b}  |
| 6. {bc a —}  | 15. {c — ab} | 24. {b a c}  |
| 7. {ab — c}  | 16. {— ab c} | 25. {b c a}  |
| 8. {ac — b}  | 17. {— ac b} | 26. {c a b}  |
| 9. {bc — a}  | 18. {— bc a} | 27. {c b a}  |

Note that " $n$  balls in 7 cells" may represent the distribution of  $n$  hits among 7 targets, or of  $n$  accidents in 7 weekdays, and so on.

Consider next the experiment of placing 3 indistinguishable balls into 3 cells. Whether or not

actual balls are indistinguishable is irrelevant; they are treated as such and, by convention, there is now a space of only 10 sample points. It is listed in tabulation (2).

- |              |             |              |
|--------------|-------------|--------------|
| 1. {*** — —} | 4. {** * —} | 8. {— ** *}  |
| 2. {— *** —} | 5. {** — *} | 9. {— * **}  |
| 3. {— — ***} | 6. {* ** —} | 10. {** * *} |
|              | 7. {* — **} |              |

In playing roulette, each point on a circle represents a possible outcome and the sample space is the interval  $0 \leq \theta < 2\pi$ . When one observes the motion of a particle under diffusion, every function  $\lambda(t)$  represents a conceivable outcome and the sample space is a complicated function space.

**Events.** In examining a bridge hand, one may ask whether it contains an ace or satisfies some other condition. In principle each such event may be described by specifying the sample points which do satisfy the stipulated condition. Thus every compound event is represented by an aggregate of sample points, and in probability theory these terms are synonymous. The standard notations of set theory are used to describe relations among events. See SET THEORY.

Given an event  $A$  one may consider the case that  $A$  does not occur. This is the negation or complement of  $A$ , denoted by  $A'$ ; it consists of those sample points that do not belong to  $A$ . Given two events  $A$  and  $B$ , the event  $C$  that either  $A$  or  $B$  or both occur is the union of  $A$  and  $B$  and denoted by  $C = A \cup B$ . In particular  $A \cup A'$  is the whole sample space  $\mathcal{S}$  which therefore represents certainty. The event  $D$ , both  $A$  and  $B$  occur, is the intersection of  $A$  and  $B$  and written  $D = A \cap B$ . It consists of the points common to  $A$  and  $B$ . If there are no such common points (as in the case of  $A$  and  $A'$ ),  $A$  and  $B$  cannot occur simultaneously and they are called mutually exclusive, written  $A \cap B = 0$ . The event " $A$  but not  $B$ " is simply  $A \cap B'$ .

**Example (c).** In tabulation (1), the event  $A$  "one cell multiply occupied" is the aggregate of the points numbered 1–21. The event  $B$  "first cell not empty" is the aggregate of the points 1, 4–15, and 22–27. Because every point belongs either to  $A$  or to  $B$  (or both),  $A \cup B = \mathcal{S}$  is the certain event. Next,  $D = A \cap B$  consists of the points 1, 4–14. Finally,  $A'$  may be described as "no cell empty."

**Probabilities in finite spaces.** If the sample space  $\mathcal{S}$  contains only  $N$  points  $E_1, \dots, E_N$  their probabilities may be any numbers such that  $P\{E_i\} \geq 0$  and  $P\{E_1\} + \dots + P\{E_N\} = 1$ . The probability  $P\{A\}$  of an event  $A$  is the sum of the probabilities of all points contained in  $A$ ; thus  $P\{\mathcal{S}\} = 1$ . To find  $P\{A \cup B\}$  one considers all points belonging to either  $A$  or  $B$ , but those belonging to both  $A$  and  $B$  are counted only once. Therefore  $P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$ . In particular, for mutually exclusive events, there is the addition rule  $P\{A \cup B\} = P\{A\} + P\{B\}$ .

Frequently considerations of symmetry lead one to consider all  $E_j$  as equally likely, that is, to set  $P(E_j) = 1/N$ . In this case  $P(A) = n/N$  where  $n$  is the number of points in  $A$ ; for a gambler betting on  $A$ , these represent the "favorable cases." For example, in throwing a pair of "perfect" dice, one naturally assumes that the 36 possible outcomes are equally likely. This model does not lose its justification or usefulness by the fact that actual dice do not live up to it, but for loaded dice a different model is required. The assumption of perfect randomness in games, card shuffling, industrial quality control, or sampling is rarely realized and its true usefulness stems from the experience that noticeable departures from the ideal scheme lead to the detection of assignable causes and thus to theoretical or experimental improvements.

How the success of probability theory depends on the disregard of preconceived philosophical ideas and on the readiness to adapt models to unexpected circumstances is illustrated by Bose-Einstein statistics.

*Example (d): Bose-Einstein statistics.* In the example of tabulation (1), the notion of perfect randomness leads to the assignment of probability  $1/27$  to each point. In the case of indistinguishable balls, tabulation (2), it has been argued that an experiment is unaffected by failure to distinguish between balls; physically there remain 27 possibilities grouped in 10 distinguishable forms. This argument leads to assigning probability  $1/27$  to each of the points 1-3, probability  $1/9$  to each of the points 4-9, and  $2/9$  to point 10. This reasoning (sound in certain situations) has been accepted as evident in statistical mechanics for the distribution of  $r$  particles in  $n$  cells (Maxwell-Boltzmann statistics). Surprisingly, it turned out that no physical particles behave this way and it was revolutionary when Bose and Einstein showed that for one type of particle all distinguishable arrangements are equally likely. This model assigns probability  $1/10$  to each point of tabulation (2). See BOLTZMANN STATISTICS; BOSE-EINSTEIN STATISTICS.

A useful, although vague, intuitive description of probability describes  $P\{A\}$  as the relative frequency of the event  $A$  if the experiment is repeated many times under identical circumstances. The laws of large numbers render this more precise, but the description often lacks operational meaning. Experiments in agriculture and human sampling cannot be repeated under remotely similar conditions, and in the case of telephone exchanges, useful probability models refer to situations which will never materialize.

**Probabilities in infinite spaces.** Two examples may illustrate the novel features of this topic.

*Example (e): unending coin tossing.* In the study of limit laws, one must consider potentially infinite sequences of coin tossings. The possible outcome of this experiment is an infinite sequence of heads and tails, and every sequence such as  $HTHTTH \dots$  represents a sample point. Finitely many tosses are the beginning of an infinite se-

quence, and the event "first four trials resulted in  $HTTH$ " is the aggregate of the infinitely many sequences with the prescribed beginning. Such an event is called an interval of length 4. There are  $2^n$  intervals of length  $n$ , and they are mutually exclusive. For reasons of symmetry, one attributes the probability  $2^{-n}$  to each interval of length  $n$ . Thus the assignment of basic probabilities refers to intervals rather than to points. A point such as  $HTHT \dots$  is the limit of an infinite sequence of contracting intervals  $H, HT, HTH, \dots$  and therefore probability zero must be attributed to each individual point.

The probabilities of other events are similarly defined by limiting procedures. For example, consider the event  $A$  that an infinitely prolonged sequence of trials never produces a run of at least two consecutive heads, or two consecutive tails. It is more convenient to enumerate the points of the complementary event  $A'$  that two equal symbols do occur in succession. Clearly  $A'$  is the union of the infinitely many mutually exclusive intervals  $HH, TT; HTT, THH; HTHH, THTT$ , and so on. Here there are 2 intervals of length  $n \geq 2$ , and therefore  $P\{A'\} = 2(2^{-2} + 2^{-3} + 2^{-4} + \dots) = 1$ , whence  $P\{A\} = 0$ . The indicated result of the experiment is thinkable, but probability zero is attributed to it. A similar, although more complicated, limiting procedure leads to the law of large numbers according to which the event "the frequencies of  $H$  and  $T$  in the first  $n$  trials tend to  $1/2$  as  $n \rightarrow \infty$ " has probability one.

*Example (f): roulette.* Here the sample space consists of the angles  $0 \leq \vartheta < 2\pi$ , and the notion of a perfect roulette assumes equal probabilities for intervals of equal length; thus an interval of length  $a$  carries probability  $a/2\pi$ . If the roulette is divided into 32 equal numbered intervals, the event "even number" consists of 16 intervals and has probability  $1/2$ .

The situation encountered here is not peculiar to probability but is common in measure theory. One starts with a collection of basic events, called intervals, and attributes probabilities to them. By simple and natural limiting procedures, probabilities can then be defined for a much wider class  $\mathcal{F}$  of events which are obtainable by applying the operations of set theory to intervals (in finite or infinite numbers).  $\mathcal{F}$  is the Borel field generated by the intervals. Probability is simply a measure on  $\mathcal{F}$ ; that is, to each event  $A$  in  $\mathcal{F}$ , there corresponds a probability  $P\{A\} \geq 0$  which is completely additive. If  $A$  is the union of the mutually exclusive events  $A_1, A_2, \dots$ , then  $P\{A\} = \sum P\{A_i\}$ . The probability of the whole space is, of course, unity.

The extension of the addition rule from finitely to infinitely many summands may be defended by considerations of continuity, but ultimately this procedure is justified by its simplicity and its success.

**Conditional probability—dependence.** Suppose that a population of  $N$  people includes  $N_A$  color-blind persons and  $N_H$  females. To the event  $A$  "a

randomly chosen person is colorblind" can be ascribed probability  $P\{A\} = N_A/N$ , and similarly for the event  $H$  that a person be female one has  $P\{B\} = N_H/N$ . If  $N_{AH}$  is the number of colorblind females, the ratio  $N_{AH}/N_H$  may be interpreted as probability that a randomly chosen female be colorblind; here the experiment "random choice in the population" is replaced by a selection from the female subpopulation. In the original experiment,  $N_{AH}/N$  is the probability of the simultaneous occurrence of both  $A$  and  $H$ , so that  $N_{AH}/N_H = P\{A \cap H\}/P\{H\}$ . Similar situations occur so frequently that it is convenient to introduce the notation

$$P\{A|H\} = \frac{P\{A \cap H\}}{P\{H\}}$$

and to call this the conditional probability of the event  $A$  relative to  $H$ . This concept is useful whenever it is desired to restrict the consideration to those cases where the event  $H$  occurs (or where the hypothesis  $H$  is fulfilled). Thus, in betting on an event  $A$  the knowledge that  $H$  occurred would induce one to replace  $P\{A\}$  by  $P\{A|H\}$ . If all sample points are equally likely,  $P\{A|H\}$  still represents the ratio of favorable cases to the total of cases possible when it is known that  $H$  has occurred.

Despite its simplicity the notion of conditional probability is exceedingly important, and frequently the probabilities in sample space are defined only in terms of conditional probabilities.

*Example (g).* In a bolt factory three machines manufacture, respectively, 25, 35, and 40% of the total. Of their output 5, 4, and 2% are defective bolts. Classification of the bolts according to the number of the machine and the quality ( $d$  for defective,  $c$  for conforming) gives the six categories  $c_1, c_2, c_3, d_1, d_2$ , and  $d_3$ . A random choice of a bolt results in one of these six outcomes, but their probabilities are not given directly. Instead, the data relating to the first machine are

$$P\{c_1 \cup d_1\} = 0.25 \quad \text{and} \quad P\{d_1|c_1 \cup d_1\} = 0.25$$

It follows that  $P\{d_1\} = 0.0125$ , and similarly for the other points. This example may also serve to illustrate the reasoning following Bayes concerning the probability of causes. Supposing a bolt was found to be defective (hypothesis  $H$ ), what is the probability that it came from the first machine (cause  $A$ )? Here

$$P\{H\} = P\{d_1\} + P\{d_2\} + P\{d_3\} \\ = 0.0125 + 0.0140 + 0.0080 = 0.0345$$

$$\text{and} \quad P\{A \cap H\} = P\{d_1\} = 0.0125$$

Thus the required answer is

$$P\{A|H\} = 0.0125/0.0345 = 25/69$$

In example (b) the probability of  $H$  "ball  $a$  is in the first cell" equals  $1/3$ , and the probability of  $A$  "first cell is multiply occupied" =  $7/27$ . Now given that the ball  $a$  is in the first cell, the conditional probability that this cell is multiply occupied becomes  $5/9$ . The knowledge that  $H$  has occurred should increase

one's readiness to bet on  $A$ . By contrast, for the event  $B$  "ball  $b$  is in the second cell,"  $P\{B|H\} = 1/3 = P\{B\}$ , and so the knowledge that  $H$  has occurred gives no clue as to  $B$ . Therefore,  $B$  is said to be independent of  $H$  if  $P\{B|H\} = P\{B\}$ , that is, if

$$P\{B \cap H\} = P\{B\}P\{H\}$$

Clearly, in this case  $P\{H|B\} = P\{H\}$  so that  $H$  is also independent of  $B$ . Accordingly, two events  $B$  and  $H$  are independent of each other if the probability of their simultaneous occurrence follows the multiplication rule  $P\{B \cap H\} = P\{B\}P\{H\}$ . This notion carries over to systems of more than two events.

**Independent trials.** The intuitive frequency interpretation of probability is based on the concept of experiments repeated under identical conditions; a theoretical model for this concept can be developed.

Consider an experiment described by a sample space  $\mathfrak{S}$ ; for simplicity of language it can be assumed that  $\mathfrak{S}$  consists of finitely many sample points  $E_1, \dots, E_N$ . When the same experiment is performed twice in succession, the thinkable outcomes are the  $N^2$  pairs of sample points  $(E_1, E_1), (E_1, E_2), \dots, (E_N, E_N)$ , and these now constitute the new sample space. It is called the combinatorial product of  $\mathfrak{S}$  by itself and denoted by  $\mathfrak{S} \times \mathfrak{S}$ ; with reference to analytic geometry, one speaks of the first and second coordinate of the point  $(E_i, E_j)$ . These notions apply equally to infinite sample spaces and to products  $\mathfrak{S} \times \mathfrak{S} \times \mathfrak{S} \dots$  of more than two factors. For example, the cartesian plane of points  $(x, y)$  is the product of the real line by itself. In tossing a coin once,  $\mathfrak{S}$  contains only the points  $H$  and  $T$ ; tossing the coin  $n$  times leads to the  $n$ -tuple product  $\mathfrak{S} \times \mathfrak{S} \times \mathfrak{S} \dots \mathfrak{S}$  whose points have  $n$  coordinates and are of the form  $(HT \dots T)$ .

Probabilities must be assigned to the events in  $\mathfrak{S} \times \mathfrak{S}$ . The case of dependent trials will be treated in the next section; if the second trial is independent of the first, the probabilities in  $\mathfrak{S} \times \mathfrak{S}$  follow the productive rule  $P\{E_i, E_j\} = P\{E_i\}P\{E_j\}$ .

In the case of  $n$  tossings of a coin, this rule leads to the probability  $2^{-n}$  for each sample point in agreement with the requirement of equally likely cases. The present approach is more flexible and more general as shown by the important Bernoulli trials.

*Example (h): Bernoulli trials.* Suppose each trial results in success  $S$  or failure  $F$ , and  $P\{S\} = p$ ,  $P\{F\} = q$  where  $p + q = 1$ . (This may be considered as the model of a skew coin.) A succession of  $n$  independent trials of this kind leads to the sample space of  $n$ -tuples ( $SFFS \dots FS$ ), and the probability of such a point is the product ( $pqqp \dots qp$ ) obtained on replacing each  $S$  by  $p$  and each  $F$  by  $q$ .

This model has obvious applications to repeated observations and to gambling. Independence is an assumption to be verified experimentally. Conceivably a coin could be endowed with memory and avoid runs of more than 17 successive heads. That

the sex distribution within families resembles Bernoulli trials is purely a matter of experience. Many gamblers fully accept the independence and yet believe that they can influence fate by using "systems," for example, by skipping the game after each failure, or waiting for a run of 3 successes, and so on. The theorem on systems shows this to be a fallacy; a gambler not endowed with foresight may use any system or random choice of the times when he plays or skips the game; he remains confronted with Bernoulli trials and is exactly in the same situation as if he played at each trial. See DISTRIBUTION (PROBABILITY).

**Example (i): geometric probabilities.** In the interval  $0 < x < 1$  a point is chosen at random. This interval is the sample space  $\mathcal{S}$  and the probability of each subinterval equals its length. The sample space  $\mathcal{S} \times \mathcal{S}$  is the unit square of the  $x, y$  plane, and the probability of any figure equals its area. The event "the two successive choices result in a sum  $< 1$ " is represented by the triangle below the main diagonal and has probability  $1/2$ . The event "the greater of the two choices is  $< t$ " is represented by the square  $0 < x < t, 0 < y < t$  and has probability  $t^2$ .

**Dependent trials; Markov chains.** Many phenomena can be analyzed in terms of dependent trials. In their description adopt the convenient and picturesque terminology of urn models, which should not detract from the general nature of the schemes.

Consider an urn containing  $N$  balls, of which  $r$  are red  $R$  and  $b = N - r$  black  $B$ . Assuming perfect randomness, the probability that a randomly drawn ball be red equals  $r/N$ . If the ball is replaced and the procedure repeated, the result is Bernoulli trials with  $p = r/N$ . Without replacement, the sample space corresponding to two drawings contains four points  $RR, RB, BR$ , and  $BB$ , to which probabilities are assigned as follows: If the first ball drawn is red (probability  $r/N$ ), the conditional probabilities of  $R$  and  $B$  at the second trial become  $(r-1)/(N-1)$  and  $b/(N-1)$ . By Eq. (1), therefore,

$$\begin{aligned} P[RR] &= r(r-1)/N(N-1) \\ P[RB] &= P[BR] = rb/N(N-1) \end{aligned}$$

and

$$P[BB] = b(b-1)/N(N-1)$$

The trials may be continued and are equivalent to ordinary sampling.

A more general urn model is obtained by letting the composition of the urn vary from trial to trial. For definiteness consider the following scheme: each time a ball is drawn, it is replaced, and  $c$  balls of the color drawn and  $d$  balls of the opposite color are added to the urn. Here  $c$  and  $d$  are fixed numbers which may be negative. This scheme contains interesting special cases such as the following:

1. When  $c = d = 0$ , drawing with replacement occurs, and for  $c = -1, d = 0$ , drawing without replacement occurs. In the latter case, the process terminates after  $N$  drawings.

2. The Polya model of contagion is the special case when  $c > 0$  is fixed and  $d = 0$ . Here the drawing of either color increases the probability of the same color at subsequent trials, just as in a contagious disease each occurrence increases the probability of further occurrences. This model represents only a crude first approximation to phenomena of contagion, but it leads to comparatively simple formulas and has been applied with astonishing success to a variety of experiences from sickness insurance to baseball scores.

3. The Ehrenfest model for heat exchange considers two containers, I and II, and  $N$  particles distributed in them. A particle is chosen at random and removed from its container into the other. This scheme differs only linguistically from the urn scheme. If the particles in I are called red, and those in II black, then each trial changes the color of one ball and gives the special case  $c = -1$  and  $d = 1$ .

The probabilities of the various possible outcomes in the general scheme are obtained as above. For example,  $P\{RBR\} = r(b+d)(r+c+d)/N(N+c+d)(N+2c+2d)$ , and so on.

Markov chains represent another important scheme for dependent trials. Suppose that at each trial the possible outcomes are  $E_1, \dots, E_N$  and that whenever  $E_i$  occurs, the conditional probability of  $E_j$  at the next trial is  $p_{ij}$ , independently of what happened at the preceding trials. Here, of course,  $p_{ij} \geq 0$  and  $p_{i1} + p_{i2} + \dots + p_{iN} = 1$  for each  $i$ . The  $p_{ij}$  are called transition probabilities. The whole process is now determined if the initial probabilities,  $\pi_i$ , at the first trial are known. For example,  $P\{E_a E_b E_c\} = \pi_a p_{ab} p_{bc}$ . The probability of the event " $E_c$  at the third trial" is obtained by summation over all  $a$  and  $b$ , and so on. Markov chains, and their analog with continuous time, represent the simplest type of stochastic process. The Ehrenfest model considered above may be treated as a Markov chain by letting  $E_i$  represent the event that container I contains  $i$  particles. Then  $p_{i,i-1} = i/N$ ,  $p_{i,i+1} = (N-i)/N$ , and  $p_{ij} = 0$  for all other combinations of  $ij$ . Other examples of Markov chains are the gambler's accumulated fortune, the composition of a deck of cards under random shuffling, and random walks. Important applications are to queueing theory where one encounters also processes with more complicated aftereffects. See QUEUEING THEORY; STOCHASTIC PROCESS.

**Random variables and their distributions.** The theory of probability traces its origin to gambling, and the gambler's gain may still serve as the simplest example of a random variable. With every possible outcome (sample point) there is associated a number, namely the corresponding gain. In other words, the gain is a function on the sample space, and such functions are called random variables. (In infinite spaces the idea is the same, but a somewhat more cautious definition is in order.) With the same experiment, one may associate many random variables. As an example consider the sample space of

tabulation (1) with probability  $1/27$  for each point. A typical random variable is the number  $N$  of occupied cells; it assumes the value 1 at the three points 1–3; the value 2 at the eighteen points 4–21; and the value 3 at the six points 22–27. One says, therefore, that the probability distribution of  $N$  is given by  $P\{N = 1\} = 1/9$ ,  $P\{N = 2\} = 2/3$ ,  $P\{N = 3\} = 2/9$ . Another variable is the number  $X$  of balls in the first cell. An inspection of tabulation (1) shows that its probability distribution is given by  $P\{X = 0\} = 8/27$ ,  $P\{X = 1\} = 12/27$ ,  $P\{X = 2\} = 6/27$ ,  $P\{X = 3\} = 1/27$ . One may also consider the two variables simultaneously and find, for example, that the combination  $N = 1$ ,  $X = 0$  occurs at two points, whence  $P\{N = 1, X = 0\} = 2/27$ . The probabilities of all pairs are given by the joint probability distribution of  $N$  and  $X$  exhibited in tabulation (3). Adding the entries in the rows

$N \backslash X$	0	1	2	3	Distribution of $N$
1	2/27	0	0	1/27	3/27
2	6/27	6/27	6/27	0	18/27
3	0	6/27	0	0	6/27
Distribution of $X$	8/27	12/27	6/27	1/27	

and columns gives the distribution of  $N$  and  $X$ , respectively, and they are therefore occasionally called marginal distributions.

Example (i) may be used to illustrate the case of continuous random variables. This example considered two consecutive selections of a point in the interval  $0 < x < 1$ . Let  $S$  be the random variable denoting the sum of the two choices, and  $L$  the larger of the two. One sees that for  $0 < t < 1$  the event  $L \leq t$  has probability  $t^2$ ; thus, setting  $P\{L \leq t\} = F(t)$  gives  $F(t) = t^2$  when  $0 \leq t \leq 1$ ; for  $t < 0$  and for  $t > 1$  one has trivially  $F(t) = 0$  and  $F(t) = 1$ , respectively. This is the distribution function of  $L$ . From it can be calculated all probabilities relating to  $L$ . Similarly the events  $S \leq u$  is represented by the region in the unit square below the line  $x + y = u$ ; therefore the distribution function of  $S$ , namely,  $P\{S \leq u\} = G(u)$ , is given by  $G(u) = 0$  for  $u \leq 0$ ,  $G(u) = (1/2)u^2$  for  $0 \leq u \leq 1$ ,  $G(u) = 1 - (1/2)(2 - u)^2$  for  $1 \leq u \leq 2$ , and  $G(u) = 1$  for  $u \geq 2$ . In like manner, the joint distribution function  $P\{L \leq t, S \leq u\} = H(t, u)$  of the pair  $L, S$  can be calculated.

Every random variable  $X$  has a distribution function  $F(t) = P\{X \leq t\}$ . If  $X$  assumes only finitely many values, then  $F(t)$  is a step function. Thus, in the example  $F(t)$  assumes the values 0,  $8/27$ ,  $20/27$ ,  $26/27$ , and 1, respectively, in intervals  $t < 0$ ,  $0 \leq t < 1$ ,  $1 \leq t < 2$ ,  $2 \leq t < 3$ , and  $t \geq 3$ . In such cases the notion of distribution function is used mainly for uniformity of language. The notion is really convenient when  $F(t)$  is not only continuous but also has a derivative  $f(t) = F'(t)$ ; then  $f(t)$  is called the probability density of  $X$ . In the above example the variable  $L$  has a density defined

by  $2t$  for  $0 < t < 1$  and 0 elsewhere; the density of  $S$  is 0 for  $u < 0$ , and  $u > 2$ ; it equals  $u$  for  $0 < u < 1$ , and equals  $2 - u$  for  $1 < u < 2$ .

The notion of independence carries over: two random variables  $X$  and  $Y$  are independent if  $P\{X \leq x, Y \leq y\} = P\{X \leq x\} \cdot P\{Y \leq y\}$ . It is easily seen that for independent variables with distribution functions  $F(t)$  and  $G(t)$  the distribution function of the sum  $S = X + Y$  is given by the convolution

$$\begin{aligned} P\{S \leq u\} &= \int_{-\infty}^{+\infty} F(u - s) dG(s) \\ &= \int_{-\infty}^{+\infty} G(u - s) dF(s) \end{aligned} \quad (4)$$

In terms of densities, Eq. (4) reads

$$h(u) = \int_{-\infty}^{+\infty} f(u - s)g(s) ds = \int_{-\infty}^{+\infty} g(u - s)f(s) ds \quad (5)$$

In the random choice example, the coordinates of the points chosen are independent variables with the rectangular density  $f(s) = g(s) = 1$  for  $0 < s < 1$ . The distribution of their sum  $S$  which has been calculated above can be found also using Eq. (5).

**Expectations.** Given a random variable  $X$  one may interpret its distribution function  $F(t)$  as describing the distribution of a unit mass along the real axis such that the interval  $a < x \leq b$  carries mass  $F(b) - F(a)$ . In the case of a discrete variable assuming the values  $x_1, x_2, \dots$  with probabilities  $p_1, p_2, \dots$  the entire mass is concentrated at the points  $x_i$ ; if  $F'(x) = f(x)$  exists, it represents the ordinary mass density as defined in mechanics. The center of gravity of this mass distribution is called the expectation of  $X$ ; the usual symbol for it is  $E(X)$ , but physicists and engineers use notations such as  $\langle X \rangle$ ,  $\langle X \rangle_A$ , or  $\bar{X}$ . In the cases mentioned one has

$$E(X) = \sum p_i x_i \quad \text{and} \quad E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

In all cases,  $E(X)$  is given by the Stieltjes integral over  $x$   $dF(x)$ . (To be precise, one speaks of expectations only when the integral converges absolutely.)

Before discussing the significance of the new concept, a few frequently used definitions are appropriate. Put  $m = E(X)$ . Then  $(X - m)^2$  is, of course, a random variable. In mechanics, its expectation represents the moment of inertia of the mass distribution. In probability, it is called variance of  $X$ ; its positive root is the standard deviation. Clearly

$$\text{Var}(X) = E(X - m)^2 = E(X^2) - m^2$$

The variance is a measure of spread: it is zero only if the entire mass is concentrated at the point  $m$ , and it increases as the mass is moved away from  $m$ . In the case of two variables  $X_1$  and  $X_2$  with expectations  $m_1$  and  $m_2$  it is necessary to consider not only the two variances  $s_i^2 = E[(X_i - m_i)^2]$  but also their covariance  $\text{Cov}(X_1, X_2) = E[(X_1 - m_1)(X_2 - m_2)] = E(X_1 X_2) - m_1 m_2$ . The covari-



ance divided by  $s_1 s_2$  is called the correlation coefficient of  $X_1$  and  $X_2$ . If it vanishes,  $X_1$  and  $X_2$  are called uncorrelated. Every pair of independent variables is uncorrelated, but the converse is not true.

If  $X_1, X_2, \dots, X_n$  are random variables with expectations  $m_1, \dots, m_n$  and variances  $s_1^2, \dots, s_n^2$ , the expectation of their sum  $S_n = X_1 + \dots + X_n$  is always given by  $E(S_n) = m_1 + \dots + m_n$ ; if all the covariances of  $X_i$  and  $X_j$  vanish, then clearly  $\text{Var}(S_n) = s_1^2 + \dots + s_n^2$ .

When  $X$  represents a physical quantity, then  $X^* = (X - m)s^{-1}$  represents the same quantity measured from a different origin and in new units. In the physicist's terminology,  $X^*$  is the quantity  $X$  referred to dimensionless units. In probability,  $X^*$  is called the reduced or standardized variable.

It was once assumed that every reasonable random variable has finite expectation and variance. Modern theory refutes this assumption. Many recurrence times in important physical processes have no finite expectations. Even in the simple coin-tossing game, the number of trials up to the time when the gambler's accumulated gain first reaches a positive level has infinite expectation.

**Laws of large numbers.** To explain the meaning of the expectation and, at the same time, to justify the intuitive frequency interpretation of probability, consider a gambler who at each trial may gain the amounts  $x_1, x_2, \dots, x_n$  with probabilities  $p_1, p_2, \dots, p_n$ . The gains at the first and second trials are independent random variables  $X_1, X_2$  with the indicated distribution and the common expectation  $m = \sum p_i x_i$ . The event that an individual gain equals  $x_i$  has probability  $p_i$  and the frequency interpretation of probability leads one to expect that in a large number of  $n$  trials this event should happen approximately  $np_i$  times. If this is true, the total gain  $S_n = X_1 + X_2 + \dots + X_n$  should be approximately  $nm$ ; that is, the average gain  $(1/n)S_n$  should be close to  $m$ . The law of large numbers in its simplest form asserts this to be true. More precisely, for each  $\varepsilon > 0$  it assures one that

$$P\left\{\left|\frac{1}{n}S_n - m\right| > \varepsilon\right\} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

This law holds also when the distribution function is not discrete.

As a special case, one can obtain a frequency interpretation of probability. In fact, consider an event  $A$  with  $P\{A\} = p$  and suppose that in a sequence of independent trials a gambler receives a unit amount each time when  $A$  occurs. Then the expectation of the individual gain equals  $p$ , and  $S_n$  is the number of times the event  $A$  has occurred in  $n$  trials. It follows that

$$P\left\{\left|\frac{1}{n}S_n - p\right| > \varepsilon\right\} \rightarrow 0$$

that is, the relative frequency of the occurrence of  $A$  is likely to be close to  $p$ .

Without this theorem, probability theory would lose its intuitive foundation, but its practical value is minimal because it tells one nothing concerning the manner in which the averages  $n^{-1}S_n$  are likely to approach their limit  $m$ . In the regular case where the  $X_j$  have finite variances, the central limit theorem gives much more precise and more useful information; for example, it tells one that for large  $n$  the difference  $S - np$  is about as likely to be positive as negative, and is likely to be of the magnitude  $n^{1/2}$ . When the  $X_k$  have no finite variances, the central limit theorem fails and the sums  $S_n$  may behave oddly in spite of the law of large numbers. For example, it is possible that  $E(X_k) = 0$  but  $P\{S_n < 0\} \rightarrow 1$ . In gambling language this game is "fair," and yet the gambler is practically certain to sustain an ever-increasing loss.

There exist many generalizations of the law of large numbers and they cover also the case of variables without finite expectation, which play an increasingly important role in modern theory. See GAME THEORY; PROBABILITY IN PHYSICS; STATISTICS. [W.F.]

**Bibliography:** W. Feller, *An Introduction to Probability Theory and Its Applications*, 2d ed., vol. 1, 1957.

## Probability in physics

To the physicist the concept of probability is like an iceberg. The part of it which he uses and which is, therefore, in full view for him is but a small fraction of what is hidden in other and larger disciplines. The philosophy and mathematics of probability have become increasingly interesting and important through many current researches. See PROBABILITY.

**Bernoulli's problem.** One of the most basic problems encountered in the application of probability to physics was solved by J. Bernoulli. It concerns the probability of achieving a specified number ( $x$ ) of successes in  $n$  independent trials when the probability of success in a single trial is known. Denote by  $p$  the probability of success; let  $q = 1 - p$  be the probability of failure. The term

$$C_x^n = \frac{n!}{x!(n-x)!} \quad (1)$$

represents the well-known binomial coefficient. The probability in question is

$$w_n(x) = C_x^n p^x q^{n-x} \quad (2)$$

For example, one may consider an urn containing  $a$  black balls and  $b$  white balls;  $n$  drawings are to be made from this urn, with replacement of the drawn ball each time. The probability that  $x$  white balls shall turn up in these  $n$  drawings is required. In this example,

$$p = \frac{b}{a+b} \quad \text{and} \quad w_n(x) = C_x^n \left(\frac{b}{a+b}\right)^x \left(\frac{a}{a+b}\right)^{n-x} \quad (3)$$

Eq. (2) defines Bernoulli's distribution; because of the Newtonian binomial coefficients which it involves, it is sometimes called Newton's formula. It satisfies the following relations:

$$\sum_{x=0}^n w_n(x) = 1 \quad (4)$$

$$\bar{x} = \sum_x x w_n(x) = np \quad (5)$$

$$\sigma^2 = \sum_x (x - \bar{x})^2 w_n(x) = \bar{x}^2 - \bar{x}^2 = npq \quad (6)$$

Hence, the standard deviation  $\sigma$  equals  $\sqrt{npq}$ .

The physicist regards the emission of particles from nuclei, of photons from hot bodies, and of electrons from hot filaments as random phenomena controlled by probability laws. One of the simplest methods for testing this assumption is based upon these formulas. Suppose that a radioactive specimen emits particles for a finite period  $T$ , and that their number is  $x$ . By repeated observations on the emission of particles in a period of  $T$  seconds, one obtains a series of numbers,  $x_1, x_2, x_3$ , etc. Now imagine the period  $T$  to be subdivided into a very large number,  $n$ , of intervals, each of length  $\tau$ , so that  $n\tau = T$ . Indeed,  $\tau$  will be assumed to be so small that, at most, one particle is emitted within the interval  $\tau$ . A single observation made during that infinitesimal interval would therefore yield the result: either no emission or emission, let us say, with probabilities  $p$  or  $q$ . By hypothesis,  $n$  such observations yield  $x$  emissions. It is seen, therefore, that the problem involves an application of Bernoulli's distribution. While it is difficult to specify the values of  $p$ ,  $q$  and  $\tau$ , the result given by Eq. (6) must nevertheless be true. Since  $q$  is extremely small,  $p$  is very nearly 1. Hence, in view of Eq. (5),  $\sigma^2 = npq = nq = \bar{x}$ . This relation has been tested experimentally and found to be true in all instances.

**Approximations.** In many practical applications, the values of  $x$  and  $n$  are very large, and a direct use of Newton's formula becomes impossible because the binomial coefficients are difficult to evaluate for large  $x$  and  $n$ . The example of particle emission just discussed is a case in point. Under these circumstances, two approximations to Newton's formula are available, one derived by C. F. Gauss, the other one by S. D. Poisson and bearing his name. Gauss's law is the limiting form of Eq. (2), when both  $x$  and  $n$  are large, so large in fact that  $1/\bar{x} = 1/np$  and  $1/npq$  are both negligible. This implies that  $p$  and  $q$  are numbers not greatly different from unity. In that case,

$$\begin{aligned} \lim_{n \rightarrow \infty} w_n(x) &= \frac{1}{\sqrt{2\pi p\bar{x}}} \exp \left[ -\frac{(x - \bar{x})^2}{2p\bar{x}} \right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -(x - \bar{x})^2 / 2\sigma^2 \right] \quad (7) \end{aligned}$$

This is Gauss's law.

If  $p$  is so small that the mean  $np$  is of the order of unity in any given application, Bernoulli's distribution is then approximated by Poisson's law, which states

$$\lim_{n \rightarrow \infty} w_n(x) = \frac{(np)^x e^{-np}}{x!} \quad (8)$$

The validity of this formula has also been confirmed in many instances of particle emission.

The density fluctuations which occur in a finite volume of gas present an interesting application of Gauss's law. Suppose that a vessel of volume  $V$  contains but a single molecule. One may then select a small element of volume  $v$  and imagine successive observations to be made in order to determine whether the molecule is in this element or not. The probability of its being there, which will be denoted by  $p$ , is clearly the ratio of the small volume element to the volume of the total container,  $p = v/V$ . Now let  $n$  observations be made upon the volume  $v$ . The probability that, in  $x$  of them, the one molecule is found in  $v$ , is given by Newton's formula. If the vessel contains not one but  $n$  molecules, then the probability of finding  $x$  of them in  $v$  simultaneously is the same as that of finding the one molecule  $x$  times in  $v$  on  $n$  successive occasions. Therefore, the probability of observing  $x$  of the  $n$  molecules in  $v$  is given by

$$w_n(x) = C_n p^x q^{n-x} \quad (9)$$

This formula may be approximated by Gauss's law. Thus, on substituting

$$p = \frac{v}{V} = \frac{1}{k} \quad (10)$$

$$\text{and} \quad q = 1 - p \quad (11)$$

there results

$$w(x) dx = \frac{k}{\sqrt{2\pi n(k-1)}} \exp \left[ -\frac{(kx - n)^2}{2n(k-1)} \right] dx \quad (12)$$

The expected mean and the dispersion of this distribution must be the same as the corresponding quantities for Bernoulli's distribution. Hence,

$$\bar{x} = np = \frac{n}{k} \quad (13)$$

$$\text{and} \quad \sigma^2 = npq = \frac{n}{k^2} (k-1) = \bar{x} \left( 1 - \frac{\bar{x}}{n} \right) \quad (14)$$

If  $k$  is sufficiently large,  $\sigma^2 = \bar{x}$  approximately.

It is often convenient to introduce a quantity  $\delta$ , known as the relative fluctuation and defined by

$$\delta = \frac{x - \bar{x}}{\bar{x}} \quad (15)$$

In view of Eq. (13)

$$\delta = \frac{kx - n}{n} \quad (16)$$

In terms of  $\delta$ , the law of fluctuations takes on the

simple form

$$w(\delta) d\delta = \sqrt{\frac{\bar{x}}{2\pi q}} \exp\left[-\frac{\bar{x}}{2q} \delta^2\right] d\delta \quad (17)$$

Here  $\bar{x}$  is the mean number of molecules within  $v$ , and  $q$  is very nearly equal to 1 if  $v \ll V$ . Formula (17) is well supported by experiment.

**Theory of errors.** Perhaps the most important application of probability theory to exact science occurs in the treatment of errors. Measurements are accompanied by errors of two kinds: determinate and random. The former arise from actual mistakes, either on the part of the observer or from faulty instruments; they are not susceptible of mathematical treatment. Random errors, however, because they are numerous, small, and likely to combine in linear fashion, are subject to the laws of probability analysis.

**The Gauss error law.** Let there be  $n$  measurements of some physical quantity resulting in the numbers  $X_1, X_2, X_3, \dots, X_n$ . If the true value of the quantity (usually unknown) is denoted by  $X$ , then the errors are

$$x_1 = X_1 - X, \quad x_2 = X_2 - X, \quad \dots, \quad x_n = X_n - X \quad (18)$$

It can be shown mathematically and has been confirmed by numerous observations that the relative frequency of occurrence of an error  $x$  is represented by Bernoulli's distribution which, under the present conditions, takes on the form of Gauss's law:

$$N(x) dx = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} dx \quad (19)$$

In this formula the parameter  $h^2$  equals  $1/(2\sigma^2)$  and is called the index of precision. Clearly, the greater the value of  $h$  the narrower the distribution given by Eq. (19). The latter is often called the Gauss error law or the normal distribution of errors. In writing it, the assumption has of course been made that the numbers  $x_1, x_2$ , etc. may be replaced by a continuous distribution.

The probability that a single measurement will contain an error between the limits  $-a$  and  $+a$  is

$$\frac{h}{\sqrt{\pi}} \int_{-a}^a e^{-h^2 x^2} dx = \frac{2h}{\sqrt{\pi}} \int_0^a e^{-h^2 x^2} dx \quad (20)$$

The integral occurring here is called the error function, and is denoted by  $\text{erf}(a)$ ; it is tabulated in most textbooks that discuss the theory of errors. The factor in front of the integral has been chosen so that

$$\int_{-\infty}^{\infty} N(x) dx = 1$$

Returning to the set of measured values  $X_1, X_2, \dots, X_n$ , one may ask: which is the most probable value (the one most likely to be true) consistent with this set of numbers? The answer is given by the principle of least squares, which affirms that the most probable value is the one for which the

sum of the squared errors

$$x_1^2 + x_2^2 + \dots + x_n^2$$

is a minimum (see LEAST SQUARES, METHOD OF). From this one may prove directly that the most probable value of  $X$  is its arithmetical mean

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \quad (21)$$

**Kinds of errors.** The reliability of a set of measurements, such as the sequence  $X_1, X_2, \dots, X_n$ , is specified by certain measures called errors, but in a slightly different sense from that previously employed. Three kinds of "error" will be described: the average error  $a$ , the root mean square error  $m$ , and the probable error  $r$ . All three refer, not to a single measurement as did the quantity  $x_i$ , but to the entire distribution (19).

The average error  $a$  is the arithmetical mean of all individual errors without regard to sign

$$a = \frac{\sum |x_i|}{n} \quad (22)$$

When the averaging process is carried out with the use of (19), one obtains the equation

$$a = \int_{-\infty}^{\infty} |x| N dx = \frac{2h}{\sqrt{\pi}} \int_0^{\infty} x e^{-h^2 x^2} dx = \frac{1}{h\sqrt{\pi}} \quad (23)$$

The measure of precision  $h$  has already been defined. It is an inverse measure of the width of the Gaussian curve, but it cannot be established without ambiguity from a finite set of measurements  $X_i$ . If the most probable value of  $h$  is calculated, it turns out to be the root-mean-square error

$$m = \sqrt{\sum x_i^2 / n} \quad (24)$$

Comparison with Eq. 23 shows that

$$a = m\sqrt{2/\pi} \quad (25)$$

The quantity  $m$  is identical with what was previously called the standard deviation.

The probable error  $r$  is defined as follows: It marks that value of  $x$  which divides the area under the curve  $N(x)$  between zero and infinity into equal parts. Hence,  $r$  is an error such that a given error  $x$  has an equal chance of being greater or smaller than  $r$ . Mathematically, the value of  $r$  is found from the equation

$$\text{erf}(hr) = 1/2 \quad (26)$$

The numerical relations between  $r, m$  and  $a$  are

$$\begin{aligned} r &= 0.6745m = 0.8453a \\ m &= 1.4826r = 1.2533a \\ a &= 0.7979m = 1.1829r \end{aligned} \quad (27)$$

**Probability in statistical mechanics.** A complex physical system made up of many constituents, for example, molecules, obeys the empirical laws of thermodynamics. Statistical mechanics is that science which attempts to explain these laws by an appeal to the laws of ordinary mechanics. In doing so, it encounters problems of the following sort.

Suppose one wishes to explain why a gas exerts a pressure upon the walls of a container. Pressure is the momentum lost by the molecules per unit time as they strike a unit area of wall. The molecules are very numerous, and their momentum losses vary erratically from instant to instant of collision, and from point to point on the surface. It is necessary, therefore, that some average be taken. Should this average be over the different collisions which a single molecule experiences in time? Or should one take the average at a given time over the whole area under consideration? Clearly, there are many ways of computing the average, each involving a particular collective in the aforementioned sense and each requiring a specification of elementary probabilities.

Now the science of mechanics does not clearly dictate which collective is the proper one to be used, and there is a considerable latitude of choice. One of these collectives which has proved most successful, the ensemble of J. W. Gibbs, will be discussed briefly.

**Phase space.** The motion of a single molecule is described in terms of two variables—its position and its momentum at any given time. In the simplest case of motion along a single axis, for example, the  $x$  axis, two numbers suffice to describe the motion; they are the position  $x$  and the momentum  $p_x$ . If a plane is constructed with  $x$  and  $p_x$  laid off on two perpendicular axes within that plane, the motion of the molecule is represented by a curve in the plane, and the plane is called the phase space of the moving molecule. A molecule whose motion is in a plane requires  $x$ ,  $y$ ,  $p_x$ , and  $p_y$ , that is, four numbers, for its complete description, and these four numbers define a point in a phase space of four dimensions. Similarly, a molecule moving in three-dimensional space has a phase space of six dimensions (axes  $x$ ,  $y$ ,  $z$ ,  $p_x$ ,  $p_y$ , and  $p_z$ ) and its dynamic behavior is depicted by a curve in this six-dimensional space.

This idea can be generalized and applied to a gas containing  $N$  molecules. The phase space of the gas will have  $6N$  dimensions, and the motion of the entire gas corresponds to the trajectory of a single point in this  $6N$ -dimensional space. A single point within it describes the physical condition of the entire gas.

**Ensembles.** Here Gibbs introduced the notion of an ensemble. He imagined a great number of thermodynamic systems, all similar to the given one. If the latter be a vessel filled with gas, he imagined a very large number of similar vessels, all filled with the same quantity of the same gas. This collection of imaginary vessels is called an ensemble. Each member of the ensemble will have its fate represented by a point moving in  $6N$ -dimensional phase space, and the whole ensemble, when viewed in that space, will appear like a cloud of dust, with each individual dust particle following its own path. The density of this cloud of dust will differ from place to place and will change in time at any given place. From the laws of mechanics it may be

shown that this imaginary cloud of dust behaves like an incompressible fluid.

There is a set of conditions, however, under which the cloud will not change its density in time, even though its individual points are in motion. One such condition amounts to the existence of a special density distribution known as the canonical distribution. It has the simple form

$$D(x_1 \cdots p_n) = \text{constant} \times \exp \left[ \frac{-H(x_1 \cdots p_n)}{kT} \right]$$

where  $H$  is the energy,  $T$  the temperature of the gas, and  $k$  is Boltzmann's constant.

At this point, contact is made with the earlier consideration. All dynamical variables, such as the pressure of the preceding example, which need to be averaged in order to correspond to the observables of thermodynamics, are to be averaged over the probability distribution  $D$ . When this is done, the laws of thermodynamics follow; in that sense, Gibbs' probability distribution provides an explanation of thermodynamics.

The success of Gibbs' theory in classical mechanics is remarkable. In order to be applicable, however, to systems which follow the laws of quantum mechanics, certain modifications are necessary. See BOLTZMANN STATISTICS; QUANTUM STATISTICS; STATISTICAL MECHANICS; STATISTICS. [H.M.]

**Bibliography:** R. Carnap, *Logical Foundations of Probability*, 1950; D. ter Haar, *Elements of Statistical Mechanics*, 1954; R. B. Lindsay, *Introduction to Physical Statistics*, 1941; R. B. Lindsay and H. Margenau, *Foundations of Physics*, 1957; R. von Mises, *Wahrscheinlichkeitsrechnung*, 1931.

## Problem solving (psychology)

The voluntary ideational behavior sustained by an organism in order to attain a goal which is not immediately accessible. Nearly identical processes have been distinguished by different terms, depending upon the circumstances under which the problem solving takes place. Usually the term problem solving has been restricted to those studies which require the subject to manipulate a tool, assemble a device, or discover the significance of a detour. If, on the other hand, the solution calls for the recognition of a relationship, that is, the ideational classification of stimuli, the process has been called concept formation or discrimination learning, when observed in infrahuman animals. When the problem to be solved requires the correction of a malfunction of equipment, the process has been called trouble shooting. When the problem requires the choosing between two or more alternative courses of action, the process has been called decision making.

Comparative psychology has made numerous contributions, not only in the description of the problem-solving capacities of lower organisms, but also in suggesting hypotheses as to the nature of the process in human beings. In addition to the

theoretical models offered by psychologists, numerous nonpsychologists have contributed to the literature on the nature of the problem-solving processes. Some have suggested the name of heuristics for this field.

**Research in problem solving.** Problem solving, in the restricted sense described above, has been studied in human and infrahuman animals. Some of the earliest research on animals dealt with the behavior of cats escaping from an enclosure, rats solving a double-alternation problem, and primates discovering the use of tools. Considerable controversy developed over the question of whether the behavior displayed by animals represented trial-and-error learning or insight. Although this issue probably was not settled to everyone's satisfaction, current research generally deals with other questions.

One major research program was concerned with the tool-using ability of anthropoid apes. The animals had to discover how to use a stick as a rake to pull in foodstuffs placed outside their cages. These studies focused attention on the visual perceptual conditions which affected such problem solving.

Analogous experiments indicate that perceptual variables are also effective in modifying the problem-solving abilities of human subjects. One series of studies found that the subjects, college students, were more likely to solve a construction-type problem if they had had prior experience with similar problems. This was especially true if they were permitted to view this previous construction while they worked on the new problem.

Trouble shooting of equipment has been studied particularly from the standpoint of decision making about probable sources of malfunction. In one series of studies, the subjects were given modified circuit diagrams and a display of "symptoms," the problem being to discover the malfunctioning component as efficiently as possible. There is considerable evidence that the ability to trouble shoot can be learned.

Several studies in human problem solving have dealt with a phenomenon called functional fixedness. Comparable behavior on verbal or numerical problems has been called set, or *Einstellung*. Functional fixedness is defined as the inability to see alternative uses for a tool or object. In one laboratory experiment, for example, a subject is instructed to make a construction using a matchbox and several other objects. If the subject is given each of the objects separately, he is more likely to succeed than if the objects are placed in the matchbox. In the latter condition the subject may fixate upon the container application of the box and ignore its other uses.

**Concept formation.** Concept formation, sometimes called concept identification or concept attainment, is the process whereby an organism, according to some self-generated principle, classifies as similar two or more stimuli which are objectively dissimilar. Some of the ways in which a dog and a

cat are similar, for example, would be that they are alive, have four feet, and have fur. Each would represent different conceptual classes which could be generated by an observer. Concept formation is important in the general field of problem solving because the problem to be solved may simply be the correct classification of stimuli, for example, which mushrooms are edible, or which control-knob shapes are most discriminable. A somewhat more complex application of concept identification in problem solving is seen when the problem solver must decide whether the tentative solution he has attained is actually a member of the class of acceptable solutions.

Three types of concepts can be differentiated in terms of the principle which must be generated by the observer. For conjunctive concepts all the members of a class have something in common; dogs, cats, and horses are all quadrupeds because they all have four legs. For disjunctive concepts the class membership is not determined by one or more common elements. The governing principle is that the class members have either one attribute or another or both; several different substances can produce a rash and still be chemically unrelated. For probabilistic concepts possession of an attribute is not a guarantee that the stimulus is a member of a class; being depressed may indicate a serious illness, but with low probability. A persistent depression, along with feelings of persecution, and frequent angry outbursts, more probably indicates the presence of serious illness, but still there is no certainty.

In order to make a decision concerning class membership the observer must receive information about the stimulus to be classified. Information which is necessary to make the correct classification is called relevant information. Information which is not necessary to make a correct classification is called irrelevant information. It is analogous to noise in a communication system and may or may not appear randomly. If the irrelevant information appears randomly it will be a considerable hindrance in arriving at the correct concept. As the amount of irrelevant information increases linearly, the difficulty of arriving at the correct solution increases at a positively accelerated rate since the number of possible combinations of attributes defining the concept increases exponentially.

Most research in concept formation has involved the use of pictures or geometric forms as stimuli with either a motor or verbal response. Recently a set of verbal stimuli, in the form of nouns, has been cataloged in terms of the frequency of particular sense impressions elicited by the words. This material is gaining considerable use and may provide a breakthrough in the problem of studying the relationship between problem solving and language (see *VERBAL LEARNING*).

Another kind of concept formation which has been widely studied in animals is called learning set. This process, sometimes called learning-to-

learn, is the acquisition of an over-all rule or principle which enables the organism to respond better than chance on a new problem. A monkey is presented with a food-well tray, and each of the three wells is covered by an object. Two of the objects are identical small red cubes, but the third object is different, a small green cone. Food is always placed under the odd object; hence this has been called the oddity problem. If the animal is given several trials with these particular objects, his performance will improve; he will learn that the reward is under the green cone. After a number of trials the cubes and the cone are replaced with another set of two identical objects and an odd object. If the animal has only learned to respond to the green cone in the first problem, his performance on the second problem should be no better than chance. If, on the other hand, the animal has learned the rule that the food is under the odd object, his performance should be better than chance. This description is an oversimplification since it usually takes experience with many problems before most animals will show any evidence of learning set. Since the learning-set paradigm has been demonstrated in a number of animal species, it has been proposed that it be used as a means of making interspecies comparisons of intellectual ability.

Human beings of course are capable of displaying learning set. It is quite likely that this process accounts for much of man's ability to generalize and to apply skills learned in one situation to a new problem. Unfortunately, man may also generalize and apply an inappropriate rule to a new situation and thereby take longer to solve the new problem than if he had had no rule.

**Group problem solving.** Group problem solving usually has been studied as a form of social psychology. Research in this area has been principally in three forms: (1) the members of the group are permitted free communication with each other; (2) the lines of communication are restricted to permit study of the effects of information flow; or (3) the subjects participate in a game played against other subjects or the experimenter or a machine.

There are numerous methodological problems which still have not been satisfactorily solved in the first-named and most common form of group problem solving. One fundamental question asked in this area of research is whether or not two heads are better than one. At best the present answer can only be that it depends upon the criterion for "better" and upon the type of problem to be solved. If the solution requires the assembly of a device or mechanical puzzle, it is obvious that a part of the device can be handled by only one person at a time. Present evidence suggests that a group will be superior to an individual when novel approaches are needed or when large amounts of information have to be stored, recalled, and related. However, because of the interdependencies of the members of the group, failure of one of the members can

adversely affect the performance of the entire group. If the criterion is the unit product per man-minute, the individual usually will be superior to the group.

**Individual differences.** This is the area of research in problem solving concerned with the assessment of ability differences of individuals through the use of tests. The usual omnibus intelligence tests are inadequate for research purposes in this area. Only since the application of factor analytic methods and the subsequent development of tests of primary mental abilities has it been possible to assess specific intellectual abilities. The information at present is sketchy.

A series of studies of young adults of above average intelligence indicates that at least 40 different factors are needed to account for differences in intellectual ability. A different number of factors, or entirely different factors may be necessary to account for the intellectual abilities of average adults and children. In addition to the difficulties of test development, there are the problems of choice of type of factor analysis and choice of criterion for the existence of a factor (see INTELLIGENCE).

**Heuristics.** This is the study of the mental processes involved in problem solving. The various stages in the process of solving a problem for which there is no immediate solution appear to be as follows:

Initially the individual must comprehend the nature of the problem. The gestalt psychologists argue for the importance of this first stage since they believe that the manner in which a problem is "seen" will determine, at least in part, the likelihood of attaining a successful solution.

The second stage, sometimes called the preparation stage, is devoted to obtaining information which appears to be relevant to the final solution. This is also a crucial stage since if the individual establishes an inappropriate set or bias, he may reject as irrelevant information which is in fact relevant. A bias in favor of irrelevant information is almost as undesirable. During this stage the problem solver tries to examine all of the relevant information and its implications for his problem. Discovery of interrelationships is in itself additional information which may be relevant.

The third stage, sometimes called the incubation stage, is characterized by the termination of active seeking for the desired solution and the turning of attention to other matters. It is usually described as a passive period of hopeful waiting for insight. Successful problem solving may follow a period of apparent inattention to the problem because the problem solver has avoided inappropriate information-handling biases during this so-called incubation period.

The final stage, when the problem solver achieves the solution and comprehends its significance, is called insight.

Clearly, the process of problem solving depends upon memory of information as well as imaginative



and creative handling of the information. See LEARNING THEORIES; MEMORY. [E.J.A.]

**Bibliography:** J. S. Bruner, J. J. Goodnow, and G. A. Austin, *A Study of Thinking*, 1956; K. Duncker, On problem-solving, *Psychol. Monograph* 58, no. 270, 1945; W. Köhler, *The Mentality of Apes*, 1926; M. Wertheimer, *Productive Thinking*, 1945.

## Proboscidea

An order of mammals including the elephants and their allies. Proboscideans first appeared in the Eocene as small pig-sized creatures, the *moeritheres*, and during the remainder of the Tertiary the stock underwent an extensive and very complex evolution. At one time or another, proboscideans lived on all continents except Australia. Two main lines of evolution are evident, the *dinotheres* and the *elephantoids*. The *dinotheres* were characterized by peculiarities of the skull and dentition. The lower jaw bore a pair of heavy tusks that curved downward and backward. The *elephantoids* are divided into three families. In the long-jawed mastodons, typified by *Gomphotherium*, the lower jaw was much elongated and there were tusks in both upper and lower jaws. The short-jawed mastodons, which lacked tusks in the lower jaw, include some of the best-known fossil vertebrates. The true elephants, characterized by very high cheek teeth, include the mammoths and modern elephants. See EUTHERIA; PROBOSCIDEA FOSSILS. [D.D.D.]

## Proboscidea fossils

Proboscideans probably originated in Africa and spread to all continents except Australia. Like their few living descendants, ancient proboscideans were large, plant-eating, quadrupedal mammals that walked on pillarlike legs. All except possibly the earliest known form in the Eocene deposits of Egypt, *Moeritherium*, had an elongate proboscis (the trunk). They also had upper and sometimes lower tusks that were enlarged second incisors. Fossil forms are classified chiefly by shape of the jaws and teeth, including the tusks, and by the character of enamel configuration in all the teeth.

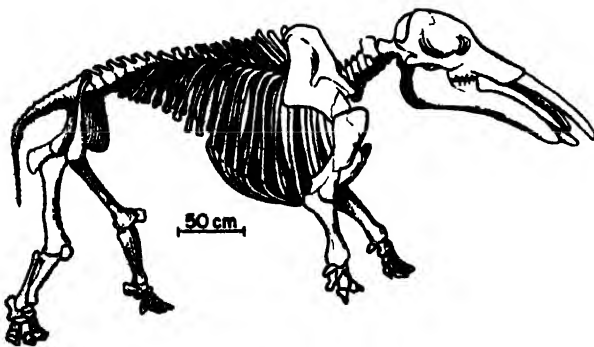


Fig. 1. Skeleton of a Pliocene long-jawed mastodont, *Gomphotherium*. (After H. Osborn, 1942)

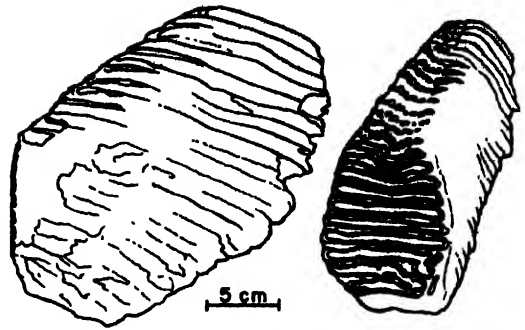


Fig. 2. Side and occlusal views of the upper third molar of a Pleistocene mammoth from Illinois. (After H. Osborn, 1942)

The tapir-sized *Moeritherium* shows many resemblances to the phenacodonts in the order Condylarthra and indicates that proboscideans probably came from phenacodonts. This evolutionary conversion probably happened in early Eocene time. See CONDYLARTHA.

Certain peculiar proboscideans, the Deinotherioidea, have the anterior part of their lower jaws abruptly down-turned. *Deinotheres* lived in Eurasia and Africa from Miocene into Pleistocene time. Another group, the Mastodontoidea (Fig. 1), includes a great array of forms of the Miocene through Pleistocene in North America, Eurasia, and Africa and the Pleistocene in South America. Mastodonts usually had low-crowned cheek teeth with tooth cusps arranged into transverse lophs (at right angles with the front-to-rear axis of the jaw). Some mastodonts (*Phiomia*, *Gomphotherium*, *Tetralophodon*, and others) had elongate lower jaws. Their lower tusks extended forward and probably served as forks to uproot clumps of vegetation upon which these animals fed. Long-jawed mastodonts like *Platybelodon* and *Amebelodon* had flattened lower tusks that may have served as shovels to scoop up masses of soft water plants. Another group of mastodonts, *Anancus*, *Stegomastodon*, *Mammut*, and others, were short-jawed and had no lower tusks. Stegodonts, generally intermediate between mastodonts and elephants, had elongate, low-crowned molars, crowned with numerous transverse crests.

The family Elephantidae (true elephants) had short lower jaws without tusks; and their high-crowned cheek teeth had the enamel, dentine, and cement arranged into transverse plates. This is interpreted as an evolutionary adaptation that enabled the teeth to withstand a greater amount of wear resulting from a diet of abrasive vegetation, such as dusty grass and tough shrubs. The elephant family includes mammoths (Fig. 2). During the Pleistocene, mammoths traveled to all the continents except Australia and South America. The famous woolly mammoth, *Mammuthus primigenius*, is well known from carcasses found in the frozen mantle rock of northern U.S.S.R. and Alaska.

Among all the living orders of mammals, elephants are most closely related probably to sea cows (*Sirenia*) and to the horse-tapir-rhino group (*Perissodactyla*). See PERISSODACTYLA FOSSILS; SIRENIA FOSSILS. [D.E.S.]

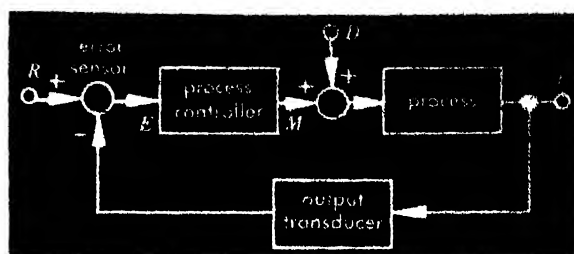
## Procellariiformes

An order of oceanic birds characterized by tube-like nostril openings, webbed feet, dense plumage, often with a peculiar musky odor, and compound horny sheath of the bill. Included families are Diomedidae (albatrosses), Procellariidae (petrels, fulmars, shearwaters), Hydrobatidae (storm petrels), and Pelecanoididae (diving petrels). The order encompasses extremes in size paralleled only by the Falconiformes, ranging from the wandering albatross (*Diomedea exulans*), with a wingspread of almost 12 ft, to the sparrow-sized least petrel (*Halocptena microsoma*). All species spend months at sea, coming to land rarely if ever, except during the breeding season. Both incubation and nestling periods tend to be exceptionally long in this order, the former being up to 80 days in the large albatrosses. Food type is related to body size, ranging from fish, squids, and other marine invertebrates, to plankton. Some species produce an oily secretion of the stomach which, together with partly digested food, is ejected as a defense mechanism. See AVES. [K.C.P.]

## Process control

A special form of feedback control system in which the element being controlled is a dynamical system referred to as a process. A process has quantities which vary with time and are related by differential equations. While the term has been associated principally with chemical and petrochemical reactions, process control is also applied to the automatic flight of missiles; the navigation of ships; the movement of machine tools; the regulation of quantities such as electric voltage, fluid pressure, temperature, reagent concentration and viscosity; and a whole variety of dynamical systems or subsystems. The essential function of a process control system is identical to that of any feedback control system, that is, the maintenance of a state of static and dynamic equilibrium between the controlled variable or variables and a reference variable or variables.

Process control systems are single-loop or multiple-loop feedback systems. The figure shows a simple, yet typical, block diagram of a single-loop, closed-loop process control configuration. The basic elements are the process being controlled, the controller, an output transducer which measures the controlled variable, and an error-sensing device. The variables in the process control system are the reference variable  $R$ , the controlled variable  $C$ , and the control error  $E$ . The output of the process controller is the manipulated variable  $M$ , which is the actuating input to the process. The function of the process controller is to modify the



Typical single-loop process control system.

control error and to produce a manipulated variable that will drive the output variable to the desired equilibrium condition relative to the reference variable.

Disturbances, which tend to upset the equilibrium condition, are introduced into the system from a number of sources. They are shown schematically in the figure as entering the system at point  $D$ . The system must minimize its response to these undesirable disturbances and remain at the equilibrium condition specified by the set point or reference variable. The specifications of the process control system may include steady-state accuracy, transient performance in response to a disturbance or a change in reference input, and low sensitivity to changes in system constants caused by environmental conditions.

The process may incorporate a number of subprocesses as part of the whole. For instance, if the over-all process is the flight of a missile, the autopilot-missile combination is the subprocess. On the other hand, if the missile is considered the process, the servomechanisms which actuate the control surfaces are subprocesses. In either case, the functions of the process or subprocess controller are governed by the same general considerations.

To have a rational design procedure for process control systems, it is necessary to obtain or estimate the differential equations that relate the manipulated (or independent) variables to the controlled (or dependent) variables. These differential equations may be linear or nonlinear. If they are linear, it is possible to express the process characteristics compactly by a transfer function, which is used to derive the required controller transfer function by the standard methods used in feedback control systems. For a discussion of the theory, see CONTROL SYSTEMS.

The more conventional application of the term process control refers to the control of a chemical process or of the many variables which affect its progress. The latter include, among others, such variables as pressure, temperature, liquid level, flow rates, reagent concentrations, and speed of rotation of mechanical elements. In a process, these quantities may be automatically controlled as subprocesses so that the main process is kept in a state of equilibrium. See CHEMICAL PROCESS CONTROL; PRESSURE CONTROL, AUTOMATIC; TEMPERATURE CONTROL, AUTOMATIC.

In evaluating or designing a process control system, consideration is given to start-up characteristics, stability, steady-state characteristics, transient response, insensitivity to disturbances, and other factors. Most processes require a start-up period, during which they progress from a state of quiescence to an active state. In a chemical process, temperature, flow, pressure, speed, valve settings, and other factors must be brought up to the condition or equilibrium that permits the chemical reaction to start. In missile control, a missile just fired from the ground requires a period of acceleration, during which its properties are rapidly changing, until it reaches a speed high enough to permit normal control procedures to be effective. Satisfying these initial requirements automatically is probably the most difficult part of process control design, because the governing relationships are time variant or nonlinear. The other design considerations are similar to those encountered in all feedback control systems and are handled in the same manner. [J.R.R.]

**Bibliography:** J. G. Truxal (ed.), *Control Engineers' Handbook*, 1958.

## Procoela

A suborder of the order Salientia characterized by a procoelous vertebral column in which each vertebra is concave anteriorly and convex posteriorly and with a free coccyx articulating by a double condyle. This suborder commonly includes five families of frogs: Leptodactylidae, Bufonidae, Atelopodidae, Hylidae, and Centrolenidae. The families Dendrobatidae and Pseudidae also are sometimes recognized.

**Leptodactylidae.** The leptodactylid frogs are most abundant in number of genera and species in the American tropics and Australia, with a very few species found as far north as the southwestern United States and New Guinea. A single genus in South Africa possibly belongs to this family. There are over 40 genera of leptodactylids with more

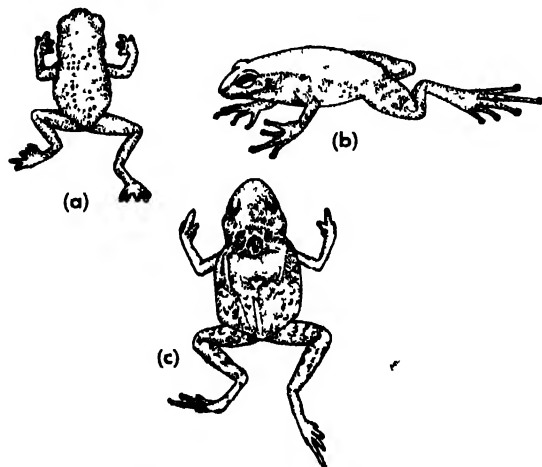


Fig. 1. (a) *Oreophrynella quelchii*, an atelopopid. (b) *Elosia nasus*, a leptodactylid. (c) *Brachycephalus ephippium*, an atelopopid. (From G. K. Noble, *The Biology of the Amphibia*, Dover, 1954)



Fig. 2. The Panamanian poison frog, *Dendrobates auratus*. (American Museum of Natural History photograph)

than 500 species; over 200 are in the genus *Eleutherodactylus* alone. Evolutionary lines within the family have undergone extensive adaptive radiation, so that one or another species lives in almost every fashion known to frogs. Species with similar habits are often similar in appearance, too, so that leptodactylid frogs are often almost indistinguishable externally from a variety of species of other families.

**Bufonidae.** Members of the large genus *Bufo*, the true toads, are native to virtually every place in the world where frogs live except the Australian region. The other four genera of the Bufonidae are found in Africa and the Malay region. These genera are of little numerical importance, but one of them, *Nectophrynoides* of East Africa, is of interest in being the only ovoviparous frog. All bufonids lack teeth, and the true toads of the genus *Bufo* are for the most part rather warty, short-legged terrestrial animals that enter the water only during the breeding season.

**Atelopodidae.** The Atelopodidae or Brachycephalidae is a relatively small family with 10 genera and about 90 species in South and Central America. Two genera in west Africa and another in the Malay region may belong in the family. Many of the atelopodids are small, brilliantly colored frogs. Members of the genus *Dendrobates* (Fig. 2) produce a skin secretion used by South American natives to poison arrows.

**Hylidae.** The Hylidae, or tree frogs (Fig. 3), is one of the larger amphibian families, with nearly 500 species known. The majority of these, about 350 species, belong to the genus *Hyla*, which is found in both eastern and western hemispheres. Many of the Hylidae are adapted to arboreal life in having expanded digital disks that facilitate climbing, but not a few have adopted other modes of existence and lead a terrestrial, aquatic, or even burrowing life.



Fig. 3. Typical North American tree frogs, the green tree frog, *Hyla cinerea* (left) and the gray tree frog, *H. versicolor* (right). (American Museum of Natural History photograph)

**Centrolenidae.** In the American tropics is found a small group of three genera and about 20 species of small, translucent, arboreal frogs grouped together in the family Centrolenidae. A peculiar characteristic of some species is that the bones are green. See AMPHIBIA; SALIENTIA. [R.G.Z.]

### Procyon

Alpha Canis Minoris, is a nearby bright star, 0.3 magnitude, of spectral type F5. Located at a distance of 3.5 parsecs, Procyon has an absolute magnitude of +2.7, which is slightly brighter than a normal main-sequence F5 star. Thus Procyon may be an old star that has begun to evolve off the main sequence. Its great age is indicated by a close faint companion of absolute magnitude +13, yellowish in color, that is, a white dwarf star, which is nearing the end of its evolution. The visual binary orbit, of 40-year period, gives Procyon a mass of 1.76, and its companion 0.65, that of the Sun. See STAR. [J.L.G.R.]

### Product design

Product design is the determination and specification of the parts of a product and their interrelationship so that they become a unified whole. The design must satisfy a broad array of requirements in a condition of balanced effectiveness. A product is designed to perform a particular function or set of functions effectively and reliably, to be economically manufacturable, to be profitably salable, to suit the purposes and the attitudes of the consumer, and to be durable, safe, and economical to operate. For instance, the design must take into consideration the particular manufacturing facilities, available materials, know-how, and economic resources of the manufacturer. The product may need to be packaged; usually it will also need to be shipped so that it should be light in weight and sturdy of construction. The product should appear significant, effective, compatible with the culture, and appear to be worth more than the price. The emphasis may differ with the instance. Durability in a

paper napkin is different from durability in a power shovel.

To determine whether a design is well adjusted to the gross array, criteria are needed. Some are objective and measurable such as clearances and efficiency, while others are quite subtle and even subjective. In a way, product design is an industrial art.

Ultimately the purpose of product design is to ascertain that the product will satisfy human wants and wishes either directly as consumer goods, or indirectly as capital equipment or components.

Except in the case of basic inventions, product design is a redesign to suit changed conditions, criteria, or enlightenment. Change may appear capricious, as in some fashions and toys, may be the result of technological progress, or may result from a change in attitude toward the product or its function. In some areas trends can be discovered, but future preferences of buyers must be predicted, gambled on, or forced to occur.

There are various steps in product design which are not necessarily in particular order. They are analytical studies; creative synthesis; drawings and models for appearance, function and specifications, plus calculations, experiments, and tests. See PRODUCTION ENGINEERING. [R.I.F.]

### Product of inertia

The product of inertia of area  $A$  relative to the indicated  $\bar{X}\bar{Y}$  rectangular axes is  $I_{\bar{X}\bar{Y}} = \int \bar{x}\bar{y} dA$  (Fig. 1). The product of inertia of the mass contained in volume  $V$  relative to the  $XY$  axes is  $I_{XY} = \int xyz \rho dV$  (Fig. 2). Similarly for  $I_{YZ}$  and  $I_{ZX}$ .

The product of inertia of area, like moment of inertia, is measured in quartic length units such as  $\text{ft}^4$ ; for mass, it is measured in mass multiplying length squared units as  $\text{g-cm}^2$ . Unlike moment of inertia, the product of inertia may be positive or negative.

Relative to principal axes of inertia, the product of inertia of a figure is zero. If a figure is mirror

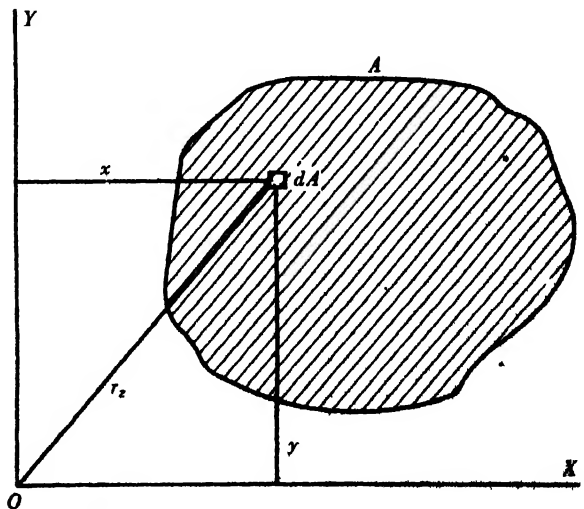


Fig. 1. Product of inertia of an area.

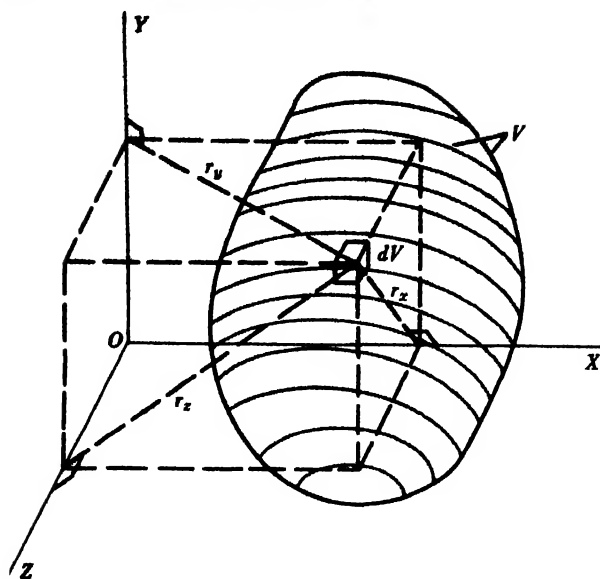


Fig. 2. Product of inertia of a volume.

symmetrical about a  $YZ$  plane,  $I_{zx} = I_{xy} = 0$ .  
See MOMENT OF INERTIA. [N.S.F.]

## Production engineering

The planning and control of the mechanical means of changing the shape, condition, and relationship of materials within industry toward greater effectiveness and value. Production engineering is a relatively new term applied to some aspects of planning and control of manufacturing; it is a service function to the production department.

As industry and technology evolve to greater levels of sophistication, complexity, and specialization the broad area of figuring out what to do becomes more involved and at the same time better understood. By this process, some of what had been originally performed by either the production department or the industrial engineer becomes a separate activity with its own background of knowledge, principles, and techniques.

**Planning and purpose.** Production engineering as a planning activity takes place between product design and the planning of the over-all manufacturing process. Over-all manufacturing planning is usually considered within the profession of industrial engineering. But in attitudes of greater specialization, production engineering may be considered a separate profession closely allied to industrial engineering.

The purpose of production engineering is to refine and adjust the design of the product (preferably with the product designer) to the problems involved in its proposed manufacture and conversely to solve certain problems, mainly mechanical, such as those involved in processing, tools, dies, and new or special equipment necessary to manufacture the product efficiently and according to the established specifications.

**Position in the organization.** Product design, production engineering, and industrial engineering

overlap variously according to the situation, policy, and organization. The techniques of production engineering are mainly in the field of mechanical engineering, but some are closely related in concept and performance to, if not directly derived from, industrial engineering.

Intelligent activity in production engineering requires a comprehensive understanding of both the intention and meaning of the product design and the means and principles of industrial engineering. The production engineer often acts as liaison between product design and industrial engineering.

The product design department specifies what is wanted, usually making only a general statement of how such specifications are to be met. The particular means are the problem of the production engineer. Product design takes into consideration performance, life, safety, and other functional requirements, usually fully testing models of various types for these features. In many products, appearance and other sensory qualities have been adjusted. These too must be maintained.

**Initial phase.** After studying the over-all product, the product engineer examines every detail of each operation for forming each piece. Various lines of inquiry are followed: Is the specific shape the most economical in material, labor, and equipment? Is it compatible with either present or obtainable equipment and know-how? Are components readily obtainable if they are to be purchased?

Usually the product designer has valid reasons for each detail, but he may not know of the benefits of alternate means, so the production engineer may often make diplomatic inquiry into any details that appear difficult, expensive, or superfluous. Prudent organizations avoid the condition in which the production engineer accepts the design specifications as absolute and final, and the opposite, where the production engineering staff can make any changes at all to suit easier processes or methods. There is usually a common-sense ground in between that can be found.

The first operations in production engineering are to examine every detail in relation to its feasibility and economy with respect to the peculiar situation. The first question most likely to arise is whether to manufacture the part or purchase it. This is not always an easy question to answer. It depends upon many factors, from over-all utilization of facilities and labor to company policy. For instance many industries have expanded, by vertical diversification, into being their own suppliers. Like much of production engineering, the answer to this lies within the policy of top management and their vision of where they are going in the light of competition, economics, technology, and cultural change. Peculiarly, here, production engineering must cooperate with sales, finance, and even basic research.

**Development of over-all production.** The first work is analytical inquiry. After most of these de-

tails are answered, if only tentatively, the major job of designing the total process begins. The whole manufacturing process usually lies within the realm of industrial engineering; production engineering deals with the mechanical aspects of manufacturing: processing equipment; tools and dies; auxiliary equipment such as fixtures, gauges, and the like; and with specially designed equipment such as conveyors, transfer equipment, automatic assembly machines, and inspection equipment. Tool design and machine design are often more specialized functions in their own right. Likewise instrumentation and control design are other special fields.

After the analytical inquiry into the detailed means, the order of performance or operation sequence is determined. This sequential study ends in a flow diagram, which is similar or supplementary to the flow-process chart and preliminary to plant layout if changes in layout are necessary.

There is a great divergence in industry. Production engineering may require nothing more than set-up and scheduling as in automatic screw machines or wire forming machines. In other instances the creation of whole new factories with special equipment may be required.

**Coordination of man and machine.** At this phase the edges of production engineering and industrial engineering diffuse with each other. Industrial engineering is concerned with methods, labor costs, and standards, but where methods become highly mechanized, dominate manual work, or are automatic, the problems become mainly ones of machine design and therefore are in the area of production engineering. From industrial engineering time data there may be indications to a mechanically oriented engineer that it would be feasible, at least economically, to search for a more mechanical means to perform certain operations. These are shifts from purely industrial engineering to an attitude toward production engineering. Many devices perform such operations as transfer, orient, differentiate, and assemble. These operations definitely sound of industrial engineering, and the industrial engineer may indicate their use, but their installation and adjustment to the total production procedure usually is a mechanical engineering problem; hence industrial engineering utilizes the mechanical bias of the production engineer.

For a long time the design, installation, use, and control of conveyors and automatic equipment have aided manufacturing, and here the mechanical production engineer and the industrial engineer have worked together. The greater the automation the more difficult it is to separate the machine designer and the production engineer from the industrial engineer in their functional position to manufacturing. It is probably safe to assign to machine design certain individual details and components and to industrial engineering the over-all plan. Production engineering then falls between and touches both.

If it is found desirable to design special equipment for production, and if the equipment is partially operated manually, principles of human engineering indicate how the operator can most effectively operate it. The equipment is in a sense an extension of the people. The force, distance, speed, accuracy, and the understanding of the operator are the limiting dimensions and factors of the equipment. The efficiencies of people in these operations determine the efficiency of the equipment. The equipment should be designed so that the operator is most efficient.

**Introduction of automation.** In newly designed equipment, especially automatic or complex machines, models are often built to find and eliminate "bugs" and to obtain data concerning time and accuracy. Often new products require new processes, which must be developed in laboratories or on prototype models. Also, unless automatic equipment is completely reliable, space should be allowed in the line for substitution of manual labor in case of a breakdown, which would be an expensive bottleneck and could close a line for days.

In industries where automation is well developed, whole lines are built as prototypes, operated to secure synchronization, to develop special skills, and to study the process for refinement of elements and their effective interaction, and for optimum application. In some industries the whole production setup is one machine and must be studied and created as a vast interacting unit. The total machine has mechanical, electrical, electronic, hydraulic, and pneumatic components operating together automatically. The production characteristics of such a machine cannot be determined from the characteristics of these separate parts. These characteristics are found by operating models often in full scale and complete in detail.

**Improvement and coordination.** The frontier of production engineering is in devising more and simpler automatic machines. The more mechanized a process becomes, the more it is freed from the limitations of human operators, and thus the more possible it becomes to mechanize it further. Because innovations reflect the individualistic approach of the engineers, several different processes or machines may be developed that do the same job equally well, that is, equally fast, efficiently, and economically.

Close control of the use of machines and detailed records of their performance and service can indicate where improvements will be most effective in increasing productivity. As technology increases in complexity, the activities that enter into it become more specialized, and the need for coordination and cooperation grows. To perform his special function efficiently, the production engineer needs to participate in conferences and planning sessions with product development engineers and factory production managers.

Forces toward obsolescence, from new materials to new viewpoints, render equipment inefficient before it wears out in the physical sense. Production



engineering strives to use advance technology wherever it provides an economic advantage. See **INDUSTRIAL ENGINEERING**; **JIG, FIXTURE, AND DIE DESIGN**; **PILOT PRODUCTION**; **PRODUCT DESIGN**; **PRODUCTION METHODS**. [R.I.F.]

## Production methods

Basically all production processes and methods can be classified as one of two types. Analytic industrial processes break down a given material into several products, as in an ore-reducing plant. Synthetic processes create one product from several different materials, as in a blast furnace or an automobile assembly plant. Processes may be a combination analytic-synthetic (wood-furniture factory) or synthetic-analytic (feed mill).

Industry frequently classifies processes into those that (1) change the shape, called forming, including cutting, molding, bending, dissolving, machining; (2) change the chemical or internal characteristics, called treating, including mixing, blending, heat treating, refining; (3) change the external surface, called finishing, including rinsing, coating, drying, painting; and (4) add other pieces, called assembling, including attaching, joining, packaging, fitting and fastening as in erecting a new building or pinning a dress pattern to a piece of cloth.

Most industries use a combination of at least two basic processes. In fact, some processes can be placed logically in more than one class, for example metal plating.

Processes have also been classified into continuous or process-type operations, as in an oil refinery, and intermittent (or repetitive) or manufacturing-type operations.

Almost all production processes or methods change the form or condition of some material, or add or deduct other materials, aided by men and/or machinery, with the end objective being a product which has greater utility by nature of its form or characteristics than the initial material.

The "end product" and the "start material" are of fundamental importance. Together, these are termed the product-material factor; this factor is the chief influencing feature in the choice of production methods. A change in the end product or in the characteristics of the start material may cause or allow significant changes in the production processes or methods.

**Design and specifications.** Design is inherent in the production of any product, even if it is not formally recognized and recorded in prints, photos, or other specifications. But generally, product designs are established prior to production and are frequently the result of several years of research, experiment, and development.

Design of products generally calls for specialists of various kinds: scientific, engineering, stylist or artistic, market research or public relations, and production engineers or manufacturing planners. The more these specialists (together with the sales, purchasing, production, and financial departments)

can integrate their views and ideas, the more effective will be the product's design for the overall company position.

Product designs may take several forms: formulae in chemical plants, blends in food products, and performance standards or drawings and specifications in manufacturing plants (see **PRODUCT DESIGN**).

The term "production design" is frequently referred to today as that design which has been engineered for ease and economy of production. It is much more than a design which merely requires functioning of the product; it involves refinements and modification of functional design based on the processes, production equipment, and personnel planned to produce the item.

The design engineer specifies what is to be made and how well it is to be made. The specifications take the form of (1) parts or materials lists describing the elements, (2) required characteristics or dimensions, (3) drawings, photos, blueprints or models, and (4) performance of finished product or test specifications. These are aimed at so describing the product and its elements that the equipment to produce it can be readily planned, the purchased materials can be correctly obtained, and the components can be made and assembled according to how the product has been engineered.

Other specifications pertain to the manufacturing process and the methods. Process specifications are set by process engineers (as distinguished from product engineers) and cover just how processes are to be controlled. Methods instruction is generally set by methods engineers and covers how work area and machinery are to be arranged.

Product specifications usually include tolerances, because nothing can be made exactly. A tolerance is a permissible variation (see **PRODUCTION ENGINEERING**).

**Standard materials and parts.** While most products require different components and even the same products require many variations, ranges and preferred choices of product dimensions or other characteristics can be established. Whenever materials and parts can be graded, classified, or otherwise standardized, great savings in time and cost result for designers, purchasers, producers and users.

For example, standard electrical current, horsepower, and dimension for electric motors allow the engineer to detail his overall machine readily, the buyer to specify and buy by numbers and code, the producer to tool up for substantial quantities of standard sizes, and the user of the machine to obtain a replacement quickly should the original motor burn out.

Standards have been established for practically all materials from lubricating oil to paper, from lumber to metallurgical specifications of sand castings. Standard gages of sheet metal, sizes of screws and nuts, and diameters of bearings are examples of standard parts we take for granted every day. These standards are set by industry associations,

the government Bureau of Standards, professional societies, or leading manufacturers (see DESIGN STANDARDS).

Standard materials and parts and standardized components lead to interchangeable manufacturing. Each component of a product is made both to fit with its mating parts and to meet a given specification, and the specification is set so that any part so made can be interchanged with the original.

Interchangeable manufacture is the underlying principle that permits mass production. By making muskets from interchangeable parts, Eli Whitney showed that a greater quantity of an article can be produced, its quality improved, and its price reduced.

Interchangeable manufacture does not necessarily result in a standardized product. Standardized parts and components can be assembled in a variety of different combinations. Current models of a popular automobile offer so many options, for example, that the manufacturer could hardly produce in the life of the model all the combinations that are numerically possible.

**Production equipment.** Machinery that actually changes the shape or characteristics of the starting material is the production equipment. Although the product design and materials generally dictate what process is to be used, the availability, suitability, and cost of the equipment to execute the process definitely affects the decision. In the final analysis, the choice of production equipment to do the job is dependent on the process selected (see MACHINING OPERATIONS; METAL FORMING).

Production equipment includes machinery, which covers the actual mechanical devices working on the product, and equipment, which covers items like paint booths, ovens, tanks, conveyors, pressure vessels, and others used directly in conjunction with operations performed on the materials. In addition, there is great variety in accessory, utility, or service equipment in any industrial facility.

Production equipment may be classed as general purpose or special. The former is universally applied to many materials or parts; the latter is designed to do one specific job, usually on one particular part.

Capacities of production equipment frequently limit a plant's operations. This capacity along with the ability to keep equipment utilized or in operation an optimum amount of time determines its productivity.

**Tools and accessories.** Smaller, easily detachable pieces of machinery or equipment used directly in production or in conjunction with the production equipment are tools and accessories. Hand tools, manual or powered, jigs, fixtures, attachments, controls, hoppers, pumps, and the like fall into this group of items. Inspection or working gages are also frequently included in the general term tools.

Just as standard parts can be assembled into a number of specific end products, so general-purpose machinery and equipment can be made to do

a special job or operation by fitting it with the proper tools or accessory equipment (see JIG, FIXTURE, AND DIE DESIGN).

**Selection of equipment.** To a great extent the selection of equipment is dictated by the production method. A degree of latitude is nearly always available within a given production method.

Much equipment, especially that not directly used in forming or assembling operations, may be common to many production methods. In selecting this equipment, consideration must be given to the following factors:

1. Demand for the product: short or long term.
2. Permanency of the product. Is it likely to remain the same or will technological changes force major changes in its design?
3. Risk of equipment obsolescence.
4. Competitive advantages or disadvantages established by choice of equipment.
5. Integration with other available or on-hand equipment.
6. Suitability of equipment in relation to the product.
7. Effect on quality of the product.
8. Quality and availability of labor to operate equipment.
9. Cost of operating and maintaining the equipment.
10. Cost, source, and availability of capital.

To resolve the first five factors requires opinions or estimates. Decisions for the second five can be obtained by evaluating facts.

A piece of equipment may be selected from standard available implements or may be built specially for the job. To decide if specially built apparatus is appropriate, additional factors must be evaluated.

1. The rate at which the product is to be made.
2. The volume or quantity of the product to be manufactured.
3. The man hours required.
4. The floor space available for the equipment.
5. The adaptability of standard machines.
6. The cost and depreciation charges of the special equipment compared with standard equipment.

To arrive at a sound decision in the selection of equipment, all phases of production under present conditions and anticipated future conditions must be carefully studied and analyzed by people experienced in the fields covered by the various factors.

Production equipment classed as general purpose does not require the extensive financial review that must be given to special equipment. The most difficult factor that must be determined for the financial appraisal of special equipment is that of product life as related to the time necessary to recover the capital invested in the equipment.

In the final analysis, the objective which must be realized in selecting production equipment is to

bring about a fair return on the money invested in that equipment. This criteria alone motivated the individuals who have foregone the opportunity of spending this money in other ways so as to provide funds for the purchase of production equipment.

**Supporting services.** In addition to the machinery, equipment, and tools used for production, every factory, plant, or mill must have certain supporting services. These take various forms.

Services dealing with materials or product include (1) production control such as planning, scheduling, machine-loading, dispatching, and recording; (2) material control, including requisitioning, receiving, storing, transporting, inventorying; (3) quality control, which includes quality levels, inspection, complaints, specifications release; (4) waste control dealing with rejects, salvage, scrap or rework; and (5) warehousing and shipping.

Services relating to machinery and equipment include (1) maintenance both preventive and repairs and overhaul; (2) tool storage and tool conditioning; (3) auxiliary or utility lines such as water, electricity, heating and ventilating, compressed air or vacuum, lubricating or cutting oil, gas, exhaust, fuel, drains, sewage, and the like.

Services relating to personnel include (1) offices; (2) restrooms, lockers, showers; (3) eating facilities; (4) parking lots and access ways; (5) time clocks, drinking fountains, first-aid, bulletin boards, telephones, and so on.

Most supporting services are necessary for a modern production facility. They must be planned into the facility and integrated with the materials, machinery, men, and the building structure. Effective arrangements and organization of these supporting facilities often account for the efficiency of the production methods established. As a result, they should not be overlooked when new or revised production methods are being planned (see PLANT FACILITIES).

[R.M.]

## Production planning

Two objectives of the manufacturing division of a business are to produce the company's products promptly and profitably. To attain these objectives requires efficient use of manpower, machines, and materials. This in turn requires coordinating the activities of each division of the company. This is done by determining the required and the available capacities and preparing a plan for matching them.

In many companies the planning is assigned to a production planning department, whose personnel also follow progress and adjust the plans to compensate for the differences between planned and actual performance. Much of this department's work is gathering, processing, distributing, and storing information.

Planning production from start to finish includes scheduling the progress of new products; forecasting demand; keeping records of available quantities of material; preparing schedules; reconciling customers' demands, inventory balances, and pro-

duction capacities; issuing realistic schedules; loading machines; dispatching; expediting; and feeding back or adjusting the next production period's schedule to compensate for the difference between planned and actual performance in the current period. Someone does each part of this work formally or informally in most manufacturing companies.

Company objectives may shift rapidly from achieving maximum output to shipping each order by the promised date to minimizing operating expenses and the inventory investment.

One essential in production planning is knowing how much time to allow between starting work and having a product ready to deliver. Some companies use a master schedule which shows the interval required between the first purchase order and the final shipment of the product. Master schedules are useful but difficult to keep up to date. Other companies prefer to use lead times they have found to be safe for each material, part, and product. See GANTT CHART; INDUSTRIAL CONTROL.

[L.F.S.]

**Bibliography:** J. F. Magee, *Production Planning and Inventory Control*, 1958; W. Voris, *Production Control: Text and Cases*, 1956.

## Progesterone

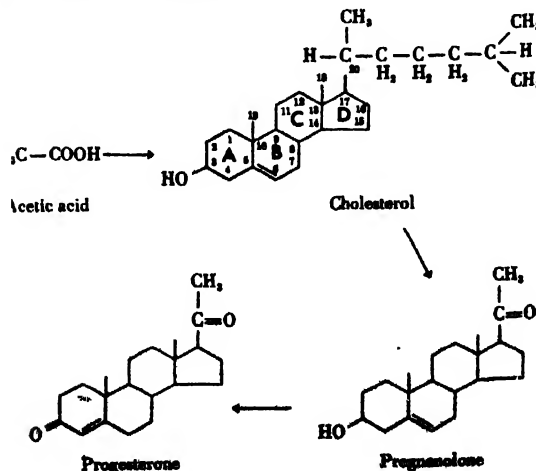
A steroid hormone produced in the corpus luteum and placenta. The hormone has an important physiological role in the luteal phase of the menstrual cycle and in the maintenance of pregnancy. In addition to these functions, progesterone produced in the testis and adrenals occupies a key role as an intermediate in the biosynthesis of androgens, estrogens, and the corticoids (adrenal cortex steroids).

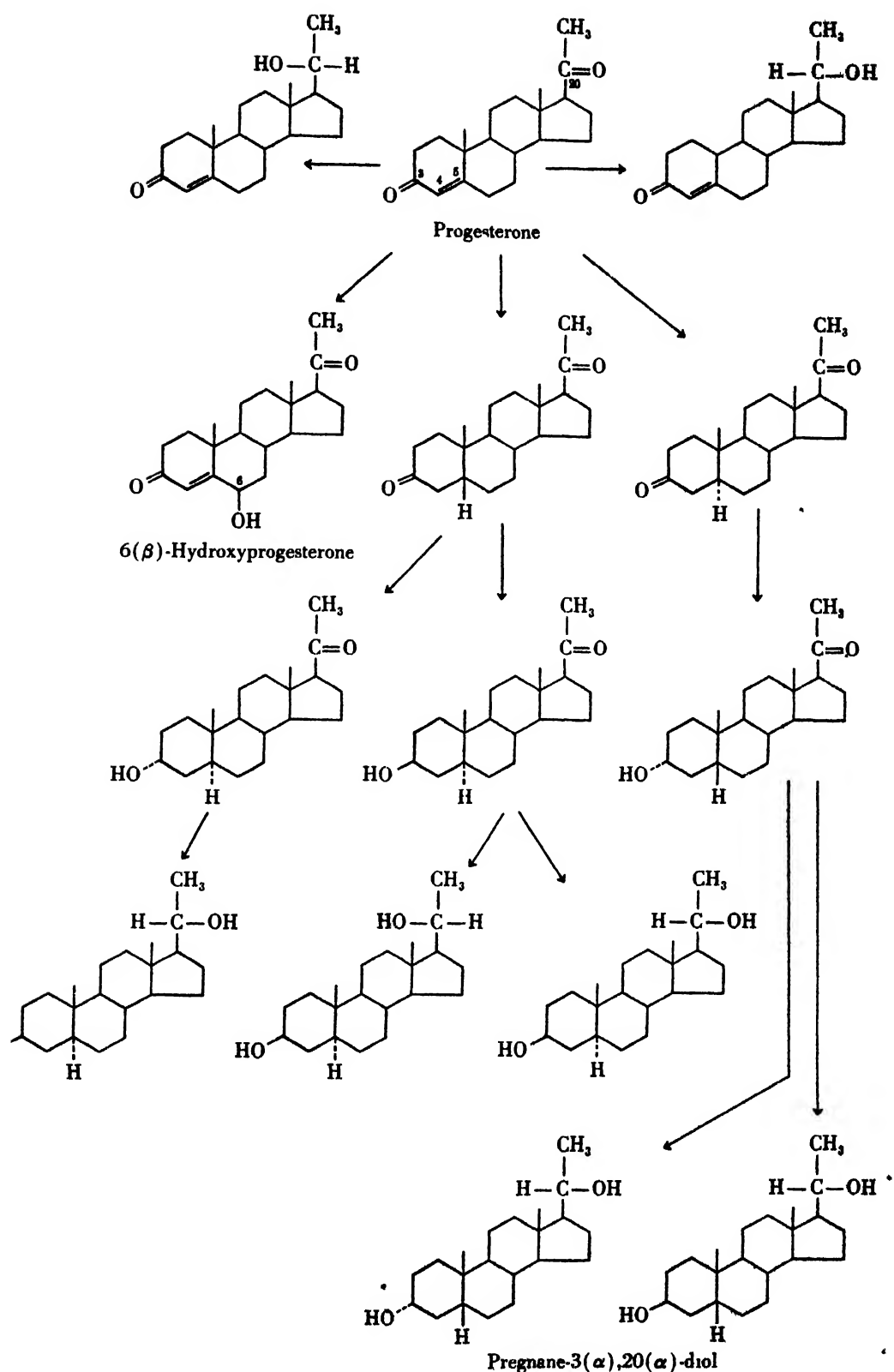
The biosynthetic pathway to progesterone is common to all of the steroid producing tissues and involves

Acetate  $\rightarrow$  Cholesterol  $\rightarrow$

Pregnenolone  $\rightarrow$  Progesterone

The blood contains, in addition to progesterone, the 20( $\alpha$ )- and 20( $\beta$ )-reduced metabolites, which have reduced biological activity.





Catabolism of progesterone involves reductive reactions at carbon atoms 3, 4, 5, and 20, as well as hydroxylation carbon atom 6 (as illustrated).

Pregnane-3(α),20(α)-diol is, quantitatively, the most important metabolite in humans, and its de-

termination in the urine is of clinical importance in conditions involving menstrual irregularities and also in abnormal pregnancies. See ANDROGEN; ESTROGEN; STEROID; STEROL.

[B.E.P.]

## Progression (mathematics)

Ordered, countable sets of numbers,  $x_1, x_2, x_3, \dots$ , not necessarily all different. In general such sets are called sequences, whereas the term progression is usually confined to the special types: the arithmetic, in which the difference  $x_k - x_{k-1}$  between successive terms is constant; the geometric, in which the ratio  $x_k/x_{k-1}$  is constant; and the harmonic, in which the reciprocals of the terms are in arithmetic progression.

**Arithmetic progressions.** If the first term is  $a$  and the common difference  $b$ ,

$$\begin{aligned} x_1 &= a, x_2 = a + b, x_3 = a + 2b, \dots, \\ x_n &= a + (n-1)b, \dots \end{aligned} \quad (1)$$

In the sum of  $n$  terms  $S_n$ , two terms equidistant from the ends always have the same sum  $x_1 + x_n$ ; hence  $2S_n = n(x_1 + x_n)$  and

$$S_n = n \frac{x_1 + x_n}{2} = n \left( a + \frac{n-1}{2} b \right)$$

If  $x_1, x_2, x_3$  are in arithmetic progression,  $x_2 = (x_1 + x_3)/2$  is called the arithmetic mean of  $x_1$  and  $x_3$ .

**Geometric progressions.** If the first term is  $a$ , the common ratio  $r$ ,

$$\begin{aligned} x_1 &= a, x_2 = ar, x_3 = ar^2, \dots, \\ x_n &= ar^{n-1}, \dots \end{aligned} \quad (2)$$

Excluding the case  $r = 1$  (when all terms are the same), the sum  $S_n$  of  $n$  terms satisfies  $S_n - rS_n = a - ar^n$ ; hence

$$S_n = a \frac{1 - r^n}{1 - r}$$

If  $|r| < 1$ ,  $r^n \rightarrow 0$  as  $n \rightarrow \infty$ ; hence the sum of the infinite geometric series

$$\sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r} \quad |r| < 1$$

If  $x_1, x_2, x_3$  are in geometric progression,  $x_2 = \sqrt{x_1 x_3}$  is called the geometric mean of  $x_1$  and  $x_3$ . Since

$$(\sqrt{x_1} - \sqrt{x_3})^2 = x_1 + x_3 - 2\sqrt{x_1 x_3} \geq 0$$

and 
$$\frac{x_1 + x_3}{2} \geq \sqrt{x_1 x_3}$$

the arithmetic mean of two unequal positive numbers exceeds their geometric mean.

The arithmetic mean  $A$  and the geometric mean  $G$  of  $n$  positive numbers are defined as

$$A = \frac{x_1 + x_2 + \dots + x_n}{n} \quad \text{and} \quad G = \sqrt[n]{x_1 x_2 \dots x_n}$$

also  $A \geq G$ .

**Harmonic progression.** The reciprocals of the sequence (1) form a harmonic progression. There is no compact expression for the sum of  $n$  terms.

If  $x_1, x_2, x_3$  are in harmonic progression,

$$x_2 = \frac{2x_1 x_3}{x_1 + x_3}$$

is called their harmonic mean.

**Sum sequence.** A general method of summing a sequence of  $n$  terms depends upon a theorem in the difference calculus which is the analog of the fundamental theorem of the differential calculus.

If  $x_n = f(n)$  is defined for  $n = 0, 1, 2, \dots$ , the difference  $f(n)$  is defined as

$$\Delta f(n) = f(n+1) - f(n)$$

$\Delta$  is a linear operator; that is,

$$\Delta[af(n) + bg(n)] = a\Delta f(n) + b\Delta g(n)$$

In the difference calculus the factorial powers  $n^{(k)}$  for integral  $k$  are defined by

$$n^{(k)} = \begin{cases} n!/(n-k)! & k < n \\ n! & k = n \\ 0 & k > n \end{cases}$$

Thus

$$\begin{aligned} n^{(k)} &= n(n-1)(n-2) \dots (n-k+1) \quad 0 < k \leq n \\ n^{(-k)} &= \frac{1}{(n+1)(n+2) \dots (n+k)} \quad k > 0 \end{aligned}$$

In particular

$$n^{(0)} = 1, n^{(1)} = n, n^{(-1)} = 1/(n+1)$$

Moreover,  $n^{(k)}$  satisfies the functional equation

$$n^{(k)} \cdot (n-k)^{(h)} = n^{(h)} \cdot (n-h)^{(k)} = n^{(h+k)}$$

The last two entries of Table 1 follow from

$$\Delta e^{in\alpha} = (e^{i\alpha} - 1)e^{in\alpha} = 2i \sin \frac{1}{2}\alpha e^{i(n+1/2)\alpha}$$

on taking real and imaginary parts.

If  $\Delta F(n) = f(n)$ ,  $F(n)$  is called the antidifference of  $f(n)$  and written  $\Delta^{-1}f(n)$ . Two antidifferences of  $f(n)$  differ at most by a constant (or by a periodic function of period 1 for functions  $f(x)$  of a continuous variable). Table 1 implies the table of antidifferences (Table 2).

Just as antiderivatives are used to compute definite integrals, antidifferences are used to compute definite sums:

$$\sum_{n=p}^q f(n) = \Delta^{-1}f(n) \Big|_p^{q+1} = F(q+1) - F(p)$$

The proof is immediate on replacing  $f(n)$  by

**Table 1. Differences**

$f(n)$	$\Delta f(n)$
constant	0
$r^n$	$(r-1)r^n$
$n^{(k)}$	$kn^{(k-1)}$
$\cos n\alpha$	$-2 \sin \frac{1}{2}\alpha \sin (n + \frac{1}{2})\alpha$
$\sin n\alpha$	$2 \sin \frac{1}{2}\alpha \cos (n + \frac{1}{2})\alpha$

$\Delta F(n) = F(n+1) - F(n)$  and performing the indicated summation.

*Example 1.* The arithmetic progression (1):

$$\sum_{n=1}^N a + b(n-1) = (a-b)n + \frac{1}{2}bn^{(2)} \Big|_1^{N+1} \\ = (a-b)N + \frac{1}{2}b(N+1)N$$

*Example 2.* The geometric progression (2):

$$\sum_{n=1}^N ar^{n-1} = \frac{ar^{n-1}}{r-1} \Big|_1^{N+1} = a \frac{r^N - 1}{r-1}$$

*Example 3.* The cosine sequence:

$$\sum_{n=1}^N \cos n\alpha = \frac{\sin(n - \frac{1}{2})\alpha}{2 \sin \frac{1}{2}\alpha} \Big|_1^{N+1} \\ = \frac{\sin(N + \frac{1}{2})\alpha - \sin \frac{1}{2}\alpha}{2 \sin \frac{1}{2}\alpha}$$

In order to sum a polynomial  $P(n)$  of degree  $k$ ,  $P(n)$  may be expressed in terms of factorial powers  $n^{(1)}, n^{(2)}, \dots, n^{(k)}$  by Newton's theorem:

$$P(n) = P(0) + \frac{\Delta P(0)}{1!} n^{(1)} + \frac{\Delta^2 P(0)}{2!} n^{(2)} \\ + \dots + \frac{\Delta^k P(0)}{k!} n^{(k)}$$

The differences of  $P(n)$  when  $n = 0$  are computed as in Table 3.

Table 2. Antidifferences

$f(n)$	$\Delta^{-1}f(n)$
0	constant
$r^n$	$\frac{r^n}{r-1} \quad (r \neq 1)$
$\begin{cases} n^{(k)} \\ 1 \end{cases}$	$\begin{cases} \frac{n^{(k+1)}}{k+1} & (k \neq -1) \\ n & (k = 0) \end{cases}$
$\cos n\alpha$	$\frac{\sin(n - \frac{1}{2})\alpha}{2 \sin \frac{1}{2}\alpha}$
$\sin n\alpha$	$-\frac{\cos(n - \frac{1}{2})\alpha}{2 \sin \frac{1}{2}\alpha}$

Table 3. Sequence of cubes

$n$	0	1	2	3	
$P(n)$	0	1	8	27	$P(0) = 0$
$\Delta P(n)$	1	7	19		$\Delta P(0) = 1$
$\Delta^2 P(n)$	6	12			$\Delta^2 P(0)/2! = 3$
$\Delta^3 P(n)$	6				$\Delta^3 P(0)/3! = 1$

In Table 3  $P(n) = n^3$  and hence

$$n^3 = n^{(1)} + 3n^{(2)} + n^{(3)} \\ \sum_{n=1}^N n^3 = \frac{1}{2}n^{(2)} + n^{(3)} + \frac{1}{4}n^{(4)} \Big|_1^{N+1} \\ = \frac{1}{4}n^2(n-1)^2 \Big|_1^{N+1} = \frac{1}{4}(N+1)^2 N^2$$

Summation by parts,

$$\Delta^{-1}[f(n) \Delta g(n)] = f(n)g(n) - \Delta^{-1}[g(n+1) \Delta f(n)]$$

is frequently useful in finding antidifferences. With  $g(n) = (-1)^n$ ,  $\Delta g(n) = 2(-1)^{n-1}$ , this gives

$$2\Delta^{-1}[(-1)^{n-1}f(n)] \\ = (-1)^n f(n) - \Delta^{-1}[(-1)^{n-1} \Delta f(n)]$$

When  $P(n)$  is a polynomial of degree  $k$ ,  $\Delta^{k+1}P(n) = 0$  and this formula repeatedly applied will give

$$\Delta^{-1}[(-1)^{n-1}P(n)]$$

and hence the sum  $\sum_{n=1}^N (-1)^{n-1}P(n)$ . The following table permits the summation of alternating powers  $(-1)^{n-1}n^k$ :

$k$	$\Delta^{-1}[(-1)^{n-1}n^k]$
1	$\frac{1}{2}(-1)^n(n - \frac{1}{2})$
2	$\frac{1}{2}(-1)^n n(n-1)$
3	$\frac{1}{2}(-1)^n(n - \frac{1}{2})(n^2 - n - \frac{1}{2})$
4	$\frac{1}{2}(-1)^n n(n-1)(n^2 - n - 1)$
5	$\frac{1}{2}(-1)^n(n - \frac{1}{2})(n^3 - n - 1)^1$

Summation by parts also gives antidifferences such as  $\Delta^{-1}[r^n P(n)]$ . Thus  $\Delta^{-1}(nr^n) = r^{n-1}p(n-p)$  where  $p = r/(r-1)$ . See SERIES. [L.B.R.]

*Bibliography:* L. M. Milne-Thompson, *The Calculus of Finite Differences*, 1933.